# DOCUMENT RESUME

ED 106 108 SE 018 869

| | |
|---|---|
| AUTHOR | Weimer, Richard Charles |
| TITLE | A Critical Analysis of the Discovery Versus Expository Research Studies Investigating Retention or Transfer Within the Areas of Science, Mathematics, Vocational Education, Language, and Geography from 1908 to the Present. |
| PUB DATE | 74 |
| NOTE | 292p.; Ph.D. Dissertation, University of Illinois at Urbana Champaign |
| EDRS PRICE | MF-$0.76 HC-$14.59 PLUS POSTAGE |
| DESCRIPTORS | *Discovery Learning; Doctoral Theses; Effective Teaching; Geography; *Historical Reviews; *Instruction; Language Instruction; *Mathematics Education; *Research; Research Reviews (Publications); Retention; Science Education; Secondary Education; Teaching Methods; Transfer of Training; Vocational Education |

ABSTRACT

In this dissertation the author has surveyed reports of experimental studies comparing discovery-oriented instruction with expository instruction. The studies analyzed focused on retention or transfer and were concerned with five subject matter areas: mathematics, science, language, geography, and vocational education. The author reports that no clear evidence of a single superior method of teaching was indicated, but, rather, that many effective teaching strategies are available. Effective strategies can be based on discovery or expository approaches and are generally usable over a broad range of content fields. Observing that several of the studies reviewed had employed questionable research designs, statistical techniques, or instrumentation, the author provides guidelines for strengthening future research in the area. He also suggests questions and problems which emerge from his study and which might be the subject of future research. (SD)

ED106108

# A CRITICAL ANALYSIS OF THE DISCOVERY VERSUS EXPOSITORY RESEARCH STUDIES INVESTIGATING RETENTION OR TRANSFER WITHIN THE AREAS OF SCIENCE, MATHEMATICS, VOCATIONAL EDUCATION, LANGUAGE, AND GEOGRAPHY FROM 1908 TO THE PRESENT

U.S. DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

Richard Charles Weimer, Ph.D.
Department of Education
University of Illinois at Urbana-Champaign, 1974

The purposes of this study were to review, analyze, evaluate, and synthesize the discovery versus expository research studies investigating retention or transfer within the areas of science, mathematics, industrial or vocational education, geography, and language. By using broad definitions of expository teaching strategy and discovery teaching strategy, experimental research studies were chosen for investigation. In addition, only those studies that employed school-related subject matter tasks and investigated retention or transfer of training were considered. The study concerned itself with identifying conceptual issues of discovery, semantic inconsistencies between studies, methodological problems in experimental designs, and problems in statistical analyses of the data; and summarizing the findings of interest.

Based on the evaluation and synthesis of the research literature on discovery, proposals are made for improving the quality of experimental research in this area and

SE 018 869

education in general for the future. Possible research problems and hypotheses are also identified.

The study is organized into five main divisions:

1. A brief history of the development and use of discovery methodology in the classroom.
2. An examination of the issues and problems dealing with the research on discovery teaching and learning.
3. A review, analysis, and summary of the comparative (discovery versus expository) research studies investigating retention or transfer in the areas of science, mathematics, industrial or vocational education, language, and geography.
4. Recommendations for improving future experimental research in methodology studies.
5. Problems and hypotheses for future research.

The review of the research literature is summarized separately for science, mathematics, industrial or vocational education, and language and geography. For each discipline, where appropriate, a separate summary is made for retention and transfer at each of four levels: elementary (grades K-6), junior-high school (grades 7-9), high school (grades 10-12), and college. The studies are reviewed in chronological order.

Within the areas of discovery and expository, this study has failed to identify a superior teaching method with respect to transfer and retention measures. Instead, a large number of teaching strategies, both discovery and expository, are identified that are effective for teaching a variety of subject-matter content. These teaching strategies should provide guidance and ideas for both preservice and inservice teacher training.

Many semantic problems were identified that existed in the research literature on discovery. Among these, the most prevalent involved the use of *discovery*, *discovery teaching*, *discovery learning*, *control group*, *retention*, *transfer*, *concept*, *principle*, *inductive method*, and *deductive method*.

The comparative research literature on discovery also revealed a number of methodological problems in experimental designs. The more salient problems identified include the following: failure to identify the significance levels in advance, failure to report power levels, using non-random selection procedures, failure to control the Hawthorne or novelty effects, failure to adequately control the teacher variable, failure to report reliability information concerning the measuring instruments, failure to report operational definitions, failure to control for pretest sensitization effects, and the use of short-term studies.

A number of controversial procedures and questionable methodology for analyzing data were detected. Among these, the most glaring were: choice of experimental unit, violations of the ANOVA model when F is less than one, covariates affected by the treatment in ANCOVA designs, indiscriminate pooling of data, improper *post hoc* comparisons, multiple comparisons and error rates, use of inappropriate statistical models, and pre-study investigations concerning homogeneity assumptions.

# A CRITICAL ANALYSIS OF THE DISCOVERY VERSUS EXPOSITORY RESEARCH STUDIES INVESTIGATING RETENTION OR TRANSFER WITHIN THE AREAS OF SCIENCE, MATHEMATICS, VOCATIONAL EDUCATION, LANGUAGE, AND GEOGRAPHY FROM 1908 TO THE PRESENT

BY

RICHARD CHARLES WEIMER

B.S., California State College (Pennsylvania), 1960
A.M., University of Illinois, 1964

THESIS

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Education
in the Graduate College of the
University of Illinois at Urbana-Champaign, 1974

Urbana, Illinois

UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

THE GRADUATE COLLEGE

July, 1974

WE HEREBY RECOMMEND THAT THE THESIS BY

RICHARD CHARLES WEIMER

ENTITLED A CRITICAL ANALYSIS OF THE DISCOVERY VERSUS EXPOSITORY RESEARCH STUDIES INVESTIGATING RETENTION OR TRANSFER WITHIN THE AREAS OF SCIENCE, MATHEMATICS, VOCATIONAL EDUCATION, LANGUAGE, AND GEOGRAPHY FROM 1908 TO THE PRESENT

BE ACCEPTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR

THE DEGREE OF DOCTOR OF PHILOSOPHY

K. B. Henderson
Director of Thesis Research

J. Myron Atkin
Head of Department

Committee on Final Examination†

K. B. Henderson
Chairman

† Required for doctor's degree but not for master's.

D517

# ACKNOWLEDGEMENTS

The author expresses his sincere appreciation to Dr. Kenneth B. Henderson, Chairman of the Doctoral Committee, for his critical comments, valuable suggestions, and knowledgeable direction. His analytical approach to education, his logical analyses of teaching, and his views on discovery teaching and discovery learning have been an inspiration to this author. Despite his heavy work load, he could always find time to discuss problems and issues.

Appreciation is also extended to Drs. Kenneth J. Travers and R. S. Jones, other members of the committee, for their critical advice and encouragement during the study. Dr. Jones' views on the nature of discovery learning have been very helpful in defining the term and understanding some of the consequences or limitations of such a definition. Dr. Travers' advice concerning the statistical and measurement areas of the manuscript merits deep appreciation.

This author is also indebted to Dr. R. Linn for his critical reading of Chapters III and V of the manuscript. His expert comments concerning the measurement aspects of the study have proved to be of significant value.

Special thanks are also extended to Dr. Maurice M. Tatsuoka for his expert advice concerning statistical matters.

Last, but not least, sincere appreciation is expressed to the author's wife, Marlene, and children, Stephanie, Ricky, and David, for their continuing patience, understanding, and encouragement.

# TABLE OF CONTENTS

CHAPTER Page

I. INTRODUCTION . . . . . . . . . . . . . . . . . . . . . . 1

Purposes of the Study. . . . . . . . . . . . . . . . 3
Justification for Conducting the Study . . . . . . 4
Organization of the Study . . . . . . . . . . . . 5

II. HISTORY OF DISCOVERY TEACHING . . . . . . . . . . . . 7

III. ISSUES AND PROBLEMS . . . . . . . . . . . . . . . . . 13

Conceptual Issues Concerning Discovery . . . . . . 14
Additional Semantic Problems . . . . . . . . . . . 22
Methodological Problems in Experimental Designs. . 39
Problems in Data Analysis . . . . . . . . . . . . 55
Summary . . . . . . . . . . . . . . . . . . . . . . 76

IV. REVIEW, ANALYSIS, AND SUMMARY OF THE RESEARCH . . . 78

Mathematics . . . . . . . . . . . . . . . . . . . . 79
Science . . . . . . . . . . . . . . . . . . . . . .184
Industrial or Vocational Education . . . . . . . .200
Geography and Language . . . . . . . . . . . . . .221
General Summary . . . . . . . . . . . . . . . . . .231

V. RECOMMENDATIONS FOR IMPROVING FUTURE RESEARCH . . .234

Experimental Designs . . . . . . . . . . . . . . .235
Statistical Analysis . . . . . . . . . . . . . . .247
Research Reports . . . . . . . . . . . . . . . . .255

VI. PROBLEMS AND HYPOTHESES FOR FUTURE RESEARCH . . . .258

Data Analyses . . . . . . . . . . . . . . . . . . .258
Learning . . . . . . . . . . . . . . . . . . . . . .258
Teaching . . . . . . . . . . . . . . . . . . . . . .261

Page

LIST OF REFERENCES . . . . . . . . . . . . . . . . . . . . . . . . . 263
VITA . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 280

# CHAPTER I

# INTRODUCTION

Over the past five decades there has been considerable discussion and research concerning the value of discovery methods in the schools. Educators and psychologists have written a great deal of hortatory literature claiming advantages and disadvantages of using discovery procedures in the classroom. A large percentage of the research effort in this area has been concerned with comparing the relative effectiveness of expository and discovery teaching strategies in the cognitive and affective domains. The reviews of the research literature comparing expository and discovery methods made by Bittinger (1968), Craig (1969), Hermann (1969), and Kersh and Wittrock (1962) have failed to conclusively identify a superior method on such criterion measures as retention, transfer, initial achievement, attitudes, interests, critical thinking, problem solving ability, and motivation. Yet, Hermann (1969, p. 58) concluded that " . . . the results tend to favor discovery teaching methods compared to other teaching methods."

Part of the problem in synthesizing the research on the subject has been the inclusion of research studies dealing with non-school related learning tasks, such as paired-associate lists, coding problems, card problems, and word relationships. These studies are useful for formulating theories of learning; but, in testing these theories, an appropriate school setting should be used.

To a large extent, teaching is based on the assumption that what is taught will be remembered and transferred to other learning situations. To an extent, transfer is dependent on retention; knowledge must be stored before it can be transferred. The exact relationship between these two constructs is not known. Transfer of training does not always occur when it might logically be expected. A great deal of experimental research has been conducted in an attempt to identify the subject matter and the conditions of learning that facilitate retention and transfer.

This author, in a preliminary study, made a review of the comparative research done on discovery teaching and learning in mathematics from 1913 to the present. No attempt was made to evaluate the experimental designs or the data analyses. Discovery teaching strategies were found to be superior to expository strategies as judged by the number of statistically significant findings reported on both retention and transfer measures. For retention, approximately 50 per cent of the studies reviewed yielded no statistically significant results concerning retention, and of the remaining studies, the discovery treatments produced approximately three times as many statistically significant findings as the expository treatments. Concerning transfer, the results were more suggestive. Not a single study favored expository methods, and approximately 60 per cent of the studies favored discovery methods as judged by statistically significant results.

A great deal of research effort has been expended comparing discovery and expository strategies for retention and transfer of learning in the areas of science and mathematics. No systematic analysis and synthesis of this body of research has been made to date.

## Purposes of the Study

The purposes of this study are to review, analyze, evaluate, and synthesize the comparative (expository versus discovery) research studies investigating retention or transfer in the areas of science, mathematics, industrial or vocational education, language, and geography. The study will concern itself with identifying semantic inconsistencies between studies, conceptuals issues of discovery, methodological problems in experimental designs, problems in the statistical analyses of the data; and summarizing the findings of interest.

Based on a synthesis and evaluation of the research literature on discovery, proposals will be made for improving the quality of experimental research in this area for the future. Possible research problems and hypotheses will also be identified.

## Justification for Conducting the Study

There are at least three reasons why this study is needed. First a critical summary of the comparative (expository versus discovery) research studies investigating retention or transfer is needed in order to determine the status of discovery and expository methods in the areas of science, mathematics, industrial or vocational education, language, and geography. No one, to this author's knowledge, has attempted to critically analyze this body of knowledge resulting from a vast expenditure of time, energy, and finances. There are many teaching behaviors which can be classified as discovery, and there are many that can be classified as expository. Both of these classifications have certain elements in common, and by using global definitions of discovery and expository, it may be possible to assess their relative effectiveness with respect to retention and transfer of school-related subject matter.

Secondly, this study is desirable for the sake of advancing the quality of experimental research in science and mathematics education as well as education in general. Knowing where the weaknesses lie and how to strengthen them should be a step forward for experimental research in education.

Thirdly, with an increased emphasis on student autonomy, activity programs, and individualized instruction, discovery techniques should receive more emphasis in the classroom. Thus, it is desirable to identify subject matter and conditions that lend themselves to discovery and expository methods.

## Organization of the Study

This study is organized into five main divisions:

1. A brief history of the development and use of discovery methodology in the classroom.
2. An examination of the issues and problems dealing with the research on discovery teaching and learning.
3. A review, analysis, and summary of the comparative (discovery versus expository)

research studies investigating retention or transfer in the areas of mathematics, science, industrial or vocational education, language, and geography.

4. Recommendations for improving future experimental research in methodology studies.
5. Problems and hypotheses for future research.

The review of the research is summarized separately for mathematics, science, industrial or vocational education, and language and geography. For each discipline, where appropriate, a separate summary will be made for retention and transfer at each of four levels: elementary (grades K-6), junior high school (grades 7-9), high school (grades 10-12), and college. The studies are reviewed in chronological order.

# CHAPTER II

# HISTORY OF DISCOVERY TEACHING

Discovery teaching is not new. The methods have been used for at least 2,000 years, though not always called by the same name. Socrates is credited with using a discovery teaching method. The method, called the Socratic questioning method, elicited and guided the student's responses toward the attainment of some goal. Socrates felt that he was not teaching his students anything, but only asking questions to help his students recollect what they had already known. Warren Colburn (1828) is often credited with introducing inductive methods into American schools. His *Intellectual Arithmetic Upon the Inductive Method of Instruction* first appeared in 1828, and developed arithmetic by using sequences of questions and simple problems. David Page (1847), the first principal of the first normal school in New York, wrote in his *Theory and Practice of Teaching*

> There is a great satisfaction in discovering a difficult thing for one's self--and the teacher does the scholar a lasting injury who takes this pleasure from him. (p. 85)

Herbert Spencer (1860), a British scientist-philosopher, wrote in his Education: Intellectual, Moral and Physical

> Children should be led to make their own investigations and draw their own inferences. They should be told as little as possible and induced to discover as much as possible. (p. 126)

Frank and Charles McMurray (1897) described the developmental method in their The Method of the Recitation as a method which

> . . . puts the questions to the child before their answers have been presented. More than that, the child is expected to conceive these answers himself; he is systematically required to make discoveries to judge what might reasonably follow from a given situation, to put two and two together and declare the result. (p. 139)

William Bagley (1905), a member of the faculty of Teachers College, Columbia University, wrote in his The Education Process

> The pupil is not to be told but led to see . . . . Whatever the pupils gain, whatever connections he works out, must be gained with the consciousness that he, the pupil, is the active agent--that he is, in a sense at least, the discoverer. (p. 262)

Discovery teaching has also been associated with heuristic teaching. According to Jones (1970), A. W. Grube seems to have made the first use of the term heuristic in the teaching of mathematics (1842). J. W. A. Young (1906), in his The Teaching of Mathematics stated that the heuristic method

> . . . is dominated by the thought that the general attitude of the pupil is to be that of a discoverer, not that of a passive recipient of knowledge. The pupil is expected in a sense to rediscover the subject, though not without profit from the fact that the race had already discovered it. (p. 8)

Concerning heuristic teaching, Johnson (1966, p. 122) stated,

> Although the term "heuristic" is one with which many teachers are apparently unfamiliar, it has long been used by educationists to denote a teaching style based on the Socratic method.

The progressive education movement, influenced by John Dewey's philosophy of nonauthoritative and child-centered instruction, had a great influence on popularizing discovery teaching techniques. This movement expressed growing dissatisfaction with the empty formalism of instruction and the unrelatedness of the school's curriculum with the child. According to Ausubel and Robinson (1969),

> . . . The progressive educator's overreaction to these faults took the form of an exaggerated emphasis on direct, immediate, and concrete experience as a prerequisite for genuine understanding, on problem solving and inquiry, and on incidental learning and learning in natural uncontrived situations. From this type of emphasis grew activity programs and project methods and the belief in "learning for and by problem solving." (p. 479)

According to Ausubel (1961), historically, the discovery method

> . . . may also be considered, in part, a revolt against the prevailing educational psychology of our time, which is largely an eclectic hodge-podge of logically incompatible theoretical propositions superimposed upon a sterile empiricism. Perhaps the most outrageous example of this unconscionable eclecticism has been the six-decade campaign sparked by Teachers College to integrate Thorndikian connectionism and a wildly extrapolated neo-Behaviorism with the major tenets of Progressive Education. (p. 21)

Many national committees have recommended discovery methods. The National Committee on Reorganization of Secondary School Mathematics in 1918 and 1923 strongly recommended discovery approaches in their reports. The Fifteenth Yearbook (1940), a report prepared jointly by committees of the National Council of Teachers of Mathematics and the Mathematical Association of America, strongly advocated discovery techniques.

Many recent and current curriculums have been sympathetic to teaching by discovery techniques. Among these are The University of Illinois Committee on School Mathematics Project (UICSM), the Nuffield Project, the Madison Project, the American Institute of Biological Science Project, the Physical Science Study Curriculum, the AAAS Elementary Science Program, and the School Mathematics Study Group (SMSG). Max Beberman (1958, pp. 38-9), in describing the UICSM program, stated, " . . . The discovery method develops interest in mathematics and power in mathematical thinking. Because of the student's independence of rote rules and routines, it also develops a versatility in applying mathematics." The Madison Project placed emphasis on autonomous decision procedures. Davis (1967, p. 61), director of the Madison Project, has cited evidence to suggest that a feeling of genuine autonomy gives rise to greater gains in learning. Some educators feel that discovery learning fosters critical thinking and creativity. Others feel that the object of a discovery lesson is to unregiment the learning routine and provide for individual differences and flexibility.

Jerome Bruner of Harvard University is one of the leading proponents of employing discovery techniques in the classroom. He (1961) and others feel that when a child learns by discovery, among other things, he understands what he learns and is better able to transfer it to new situations, he has better retention for longer periods of time, he is motivated and interested in his work, he learns searching strategies for solving problems, and he develops more favorable attitudes towards the subject.

Historically, discovery teaching has been among the repertoire of accepted techniques available to teachers.

Much hortatory and research literature has been written to promulgate its effectiveness in the classroom.

# CHAPTER III

## ISSUES AND PROBLEMS

An examination of the literature on discovery teaching and discovery learning has revealed that there are many issues, problems, and inconsistencies. Many educators and psychologists use the term discovery loosely, and many different conceptual uses of the term exist. In experimental research literature, there exist problems in design and statistical analyses of the data. The designs often contain confounding variables that prohibit valid inferences to be drawn from the studies. Many semantic inconsistencies exist in labeling activities or defining terms used in the research literature. Furthermore, in some studies the data are analyzed by using inappropriate statistical models or questionable procedures.

The reader may find it helpful or instructive to read Chapter IV prior to reading this chapter. The present sequence was chosen to alert the reader to the issues and problems encountered in analyzing the research in Chapter IV.

Not all the studies cited in this chapter are reviewed in Chapter IV. For the most part, either these studies not reviewed did not investigate retention or transfer, or, if they did, they did not report any statistically significant findings concerning retention or transfer. Some studies reporting nonsignificant findings in the areas of retention or transfer were included in the review to reveal certain problems or to illustrate certain teaching procedures or techniques.

## Conceptual Issues Concerning Discovery

### Discovery Teaching versus Discovery Learning

Discovery is sometimes used synonymously with discovery teaching and also with discovery learning. It is doubtful that this indiscriminate usage is fruitful.

The focus of teaching is on the behavior of the teacher; the focus of learning is on the behavior of the learner. There is a distinct different between teaching and learning. Concerning this difference, Green (1971) states:

> Insofar as the activity of teaching involves giving reasons, evidence, explanations, and conclusions, it can be evaluated independently of its results in getting someone to learn. Teaching, insofar as it is limited to the logical acts, can be well done, even though nobody learns, because giving reasons, evidence, or explanations can be well done even though nobody learns from it. The performance of the logical acts of teaching is appraised on logical grounds. (pp. 7-8)

The effectiveness of teaching is often determined or measured by the change in behavior of the student. But it must be kept in mind that teaching is neither necessary nor sufficient for learning to take place; one can take place without the other. According to Green (1971),

> . . . Teaching cannot be understood as the kind of activity that causes learning, because it can occur when learning does not. Moreover, learning can occur when there is no teaching. (p. 140)

He (1971) further states:

> . . . To suppose that learning is the effect of which teaching is the cause, that learning is produced by teaching or caused by teaching, is to commit a category mistake. . . . Teaching does sometimes contribute to learning but probably not in a causal or logically necessary way. Learning is not the product of teaching. To search for the universally successful method is like trying to find the infallible way in which looking for my cuff links will always result in find [sic] them. (p. 141)

The student, being exposed to an expository lesson, may be learning by discovery methods, particularly if he is lost in the lesson. Also, the best discovery strategies will not necessarily lead all students to discover.

## Learning by Discovery

Discovery learning has been viewed from many different perspectives. Ausubel (1963, p. 16) has defined discovery learning as " . . . the principal content of what is to be learned is not given. . . ." Bruner (1961, p. 22) defines *discovery learning* as " . . . a matter of rearranging or transforming evidence in such a way that one is enabled to go beyond the evidence so reassembled to additional new insights." Ballew (1967, p. 262) defines *discovery learning* as that which " . . . means that the student assembles bits of evidence and then goes beyond this evidence to attain some knowledge, not necessarily new to mankind, but new to the student." For Zubulake (1970, p. 7), *discovery* is " . . . a method of problem solving with no central frame of reference and very little teacher guidance."

Frequently *learning by discovery* refers to an initial process in the learner and is referred to as an intervening variable. Kersh (1964, p. 227) uses *discovery* to describe a learner's goal-directed behavior when he is forced to complete a learning task with little or no help from the teacher. If the learner completes the task with little or no help, he is said to have learned by discovery. Suchman (1962) stated:

> Discovery can be thought of as the experience associated with the sudden assimilation of perceived data within the framework of a perceptual system regardless of whether this was brought about by the reorganization of the data or of the system. (p. 3)

For Gagne (1966), discovery learning

> . . . may be said to occur when the performance change that is observed requires the inference or an internal process of search and selection. (p. 149)

Price (1965), after reviewing the literature on discovery, defined *discovery* as a dichotomy between teacher and learner; the teacher creates an atmosphere of intellectual curiosity which leads the student to discover meaningful subject matter. He (p. 35) identifies five aspects of learning by discovery: initiation, delineation, speculation and investigation, generalization, and adaptation. The student may involve many of these aspects, in any order, to attain his goal of discovery. The initiation stage may involve either the teacher or the student (in the case of pure discovery).

This author proposes the following operational definition of *discovery learning*.

> *S has learned subject matter M by discovery* if there exists a time interval $(t_1,t_2)$ and a valid test T such that at time $t_1$, S does not possess M as determined by T; at time $t_2$, S possesses M as determined by T; and during the time interval $(t_1,t_2)$, S has not been communicated M by an external source.

According to this definition, if a person is to learn a certain fact, and he looks it up in an encyclopedia, he has not learned the fact by discovery. In this case, the student has been communicated the fact by an external source, the encyclopedia. Under this definition, learning aids are permitted to facilitate discovery as long as they do not verbally communicate the goals of the lesson.

## Discovery Teaching

The most widespread use of *discovery* has been with reference to the teaching method. Discovery teaching methods are frequently contrasted with methods of instruction labeled as *didactic*, *expository*, *teacher-centered*, *deductive*, *direct-detailed*, and *reception learning*.

Many different teaching methods have been identified as being discovery methods, the most popular being the inductive method. In the inductive method, the teacher (person, textbook, or machine) presents a sequence of instances or examples, depending on whether the goal of instruction is learning a principle or a concept, that the learner can manipulate, analyze, or experiment with. The pupil may or may not discover the goals of the lesson. Carroll (1964) defines *inductive discovery teaching of a concept* as

> . . . presenting an individual with an appropriate series of positive and negative instances of a concept, labeled as such, and allowing him to infer the nature of the concept by noticing invariant features or attributes. (p. 202)

Another popular discovery method for teaching a principle is the Socratic questioning method. In this method, the teacher controls the data used by the students, since his questions must elicit propositions from which the students can deductively infer the object of the lesson. The Socratic questioning method is a form of deductive-discovery teaching. In a deductive-discovery teaching strategy, the teacher attempts to operate in such a fashion that the students, after considering one or more propositions, attempt to infer the object of the lesson, usually a rule or generalization.

A study by Levine (1967), comparing two discovery methods of teaching vector geometry concepts to college students, is useful for illustrating two types of discovery teaching, inductive and deductive. The inductive method was labeled as the *experience-to-theory* approach, while the deductive method was labeled as the *theory-to-application* method. With the experience-to-theory approach,

> . . . the instructor presented an instance of a theorem as a problem (That this was an instance of a theorem was known to the instructor but not to the students). Solving this problem was the first step toward the formulation of a generalization for the students. This was followed by the instructor's presenting a second problem--also an instance of the theorem. As many additional problems as seemed necessary were presented, until the instructor was able to secure from the students a suitable general statement to prove. With the instructor's helping out when necessary, the students developed an acceptable proof for the generalization. Additional problems, representing applications of the theorem (or generalization), were presented, enabling the students to develop a means of attacking the problems and analyzing the results. (pp. 9-10)

In the theory-to-application approach,

> . . . the instructor began with a statement of a theorem, for which students proceeded to develop a proof. The members of the class were involved in questions and answers relating to the proof which was being developed at the blackboard by one of the students. During this time, the instructor remained in the background, answering questions which the students could not resolve by themselves. After the proof was completed, problems involving applications of the theorem were posed by the instructor. (pp. 8-9)

Concerning the conceptual issues of discovery teaching, Romberg and DeVault (1967) state:

> . . . The importance of the concept of discovery to curriculum is widely recognized. Research relative to the concept, however, has not been very helpful because of the lack of clarity in defining the units of behavior in the teaching act. Because the concept has been so important in mathematics education, further critical analyses of the process are valid sources of mathematics curriculum research. (p. 100)

It is important to realize that there are differences in what is called discovery-type teaching; there is no one common use of the term discovery. The meaning of discovery differs noticeably from one study to another. Such semantic inconsistencies seem to abound in educational and psychological literature.

A wide variety of terminology is used to describe discovery teaching methods. Among these, in addition to the ones mentioned above, are heuristic methods, activity learning, laboratory methods, pure discovery, the incidental method, open-ended experiments, inquiry methods, free experimental techniques, guided discovery, and example-rule strategies.

Often, if two discovery strategies differ, they differ in the amount and kinds of guidance or directedness given to the student to direct his learning. Expository (telling everything) and pure discovery (telling nothing) represent the two extremes; guided discovery falls somewhere in between. Hence, it is possible to classify discovery strategies along the continuum of guidance.

Guidance can take on many different forms. The various forms of guidance include the following: mode of presentation, giving praise, giving rules, giving answers, giving hints, giving instructions, sequencing or grouping data into patterns, varying the sequencing in a lesson, providing encouragement, pre- and post-organizers, degree of meaningfulness, use of models or other audio-visual aids, and degrees of abstractness. In most cases it has not been possible to analyze research studies along this dimension because of the lack of uniformity between studies.

### Learning by Discovery, Process or Product?

Learning by discovery can be thought of as a process or as a means to an end, or as a product or goal. For Bruner, discovery learning is a process. He (1966) stated that

> . . . a theory of instruction seeks to take account of the fact that a curriculum reflects not only the nature of knowledge itself--the specific capabilities--but also the nature of the knower and of the knowledge getting process. It is the enterprise par excellence where the line between the subject matter and the method grows necessarily indistinct. A body of knowledge, enshrined in a university faculty, and embodied in a series of authoritative volumes is the result of much prior intellectual activity. To instruct someone in these disciplines is not a matter of getting him to commit the results to mind; rather, it is to teach him to participate in the process that makes possible the establishment of knowledge. We teach a subject not to produce little living libraries from that subject, but rather to get a student to think mathematically for himself, to consider matters as a historian does, to take part in the process of knowledge-getting. Knowing is a process, not a product. (p. 72)

For Gagne and others, the focus of discovery learning is the product or end result. As Gagne (1965) stated:

> . . . Knowing a set of strategies is not all that is required for thinking; it is not even a substantial part of what is needed. To be an effective problem solver, the individual must somehow have acquired masses of structurally organized knowledge. Such knowledge is made up of content principles, not heuristic ones. (p. 170)

Furthermore, to produce the ability to discover (an end) may involve more than simple practice at discovery (a means).

## Additional Semantic Problems

In addition to the meaning of discovery, which was discussed in the last section, many other semantic problems exist in the research literature on discovery. Among these, the most prevalent involve the use of *control group*, *retention*, *transfer*, *concept*, *principle*, *inductive method*, and *deductive method*.

### Control Group

A control group is included in an experimental design for comparative purposes. On some occasions, the control group is exposed to an experimental treatment; on others, the control group receives only the criterion measures. In this later case, if there is a statistically significant difference between experimental and control group means, the difference is attributed to the treatment (considering a well controlled design) and not to the maturation of the subjects. In the former case, where the control group is exposed to a treatment, a significant difference does not necessarily indicate a difference that would be attributable to maturation factors.

These two uses of *control group* lead to much confusion and many misconceptions. For example, Price's (1965) control group was administered an expository treatment, and his two experimental groups were given discovery treatments. The experimental and control classes were both exposed to similar material; only the mode of presentation and sequencing was changed. Brudzyski (1966, p. 2) used control group to indicate the children who were involved in the lecture demonstration instructions." In a study by Schaaf (1954), the control group received instruction similar to that received by the experimental group. The control group was exposed to a "traditional treatment", and the experimental group was exposed to a discovery treatment. Schaaf's use of the control group was to control for any changes that might result from normal maturation factors. The confounding of maturation factors and treatment make it extremely difficult, if not impossible, to determine the effects of maturation. In describing his control groups, Brown (1969, p. 64) stated, "The students in the conventional classes were used as the control groups." Craik's (1966, p. 56) control groups were exposed to the deductive-descriptive method of teaching.

The majority of the experimental designs reviewed for this study which included a control group exposed the control group to both a "placebo" treatment and the criterion tests.

The use of the term control group in experimental studies often depends on the nature of the experiment. In some cases, control group refers to the treatment contrasted with the experimental treatment; and, in others, control group refers to a group used to make baseline comparisons or to determine the effects of history or maturation factors.

## Retention

In some studies it is difficult to differentiate between initial learning or achievement measures and a retention measure. In a study by Lahnston (1972), the initial acquisition test is called a retention test. McConnell's (1934) seven-month experiment with teaching basic addition and subtraction facts to second-grade students involved 14 tests plus a pupil questionnaire. Some of these tests demonstrated that certain addition and subtraction skills were retained by students. It is questionable how long these facts were retained.

Belcastro (1966) compared two programed methods (inductive and deductive) of teaching certain algebraic content to eighth-grade students. After the three-day learning program, a 28-item posttest was administered. The experimenter (p. 79) refers to this test as measuring retention when he stated, "The deductive method of programing definitely resulted in a superior average retention compared to the inductive method of programing materials."

Gagne (1970) states that learning involves retention.

> . . . Learning as a total process begins with a phase of apprehending the stimulus situation, proceeds to the phase of acquisition, then to storage, and finally to retrieval. All of these events are involved in the sequence of an act of learning. . . . (p. 78)

According to Gagne (1970), retention involves the storage and retrieval of information. Thus, initial learning or achievement, as demonstrated by student behavior, indicates retention; in fact, all achievement measures are retention measures.

A crucial interest in retention studies is the length of time between acquisition and retrieval of information. In studies with only one measure of achievement (e.g., Belcastro, 1966) this is difficult to determine. For, if certain content is taught at time $t_1$, and a retention test is administered at time $t_2$, one cannot assume that the learning was retained over the time interval $(t_1, t_2)$. It may be the case that the student learned the subject-matter content somewhere between time $t_1$ and time $t_2$. The crucial point is the time of storage.

Many experimenters (e.g., Fullerton, 1955; Werdelin, 1968; Barrish, 1971; and Bassler, et al., 1971) avoided the length of storage dilemma by administering two equivalent forms of an achievement test, one to measure storage and one to measure retrieval. The retention tests employed by Bassler et al. (1971) are characteristic of tests in studies employing equivalent or parallel forms.

> The retention test, which was administered approximately four weeks after the termination of instruction, was a shortened version of the posttest in the fourth and sixth grade studies. The eighth

> grade retention test was a parallel form of the posttest. The content of items on the retention test was unaltered; however, all retention items were changed in context from their analogues on the posttest. (p. 307)

Some studies (e.g., Norman, 1955; Anastasiow et al., 1970; Rizzuto, 1970; Cooke, 1971; Murdoch, 1971; and Olander and Robertson, 1973) employed the posttest to measure both storage and retrieval with a time interval in between, typically from one to six weeks.

The typical retention study encountered by this author was concerned with comparing the retention measure raw scores made by students exposed to a discovery treatment with the retention measure raw scores made by students exposed to an expository treatment. Rarely, if ever, did the employed measuring instruments measure what was retained by the students. With just comparing raw scores, certain items could have been retrieved on the posttest and different items retrieved on the retention or delayed posttest.

In some studies, retention was concerned with the retrieval or reinstatement of intellectual skills as well as verbal information. Grote (1960) defined retention as

> . . . the degree to which subjects were able to retain associated facts, and apply principles of mechanics in the solution of mechanical problems that were similar to, but not identical to, those studied during the instructional periods as measured one and six weeks after instruction. (p. 10)

In describing his retention test, Meconi (1967, p. 52) stated that " . . . the test items were constructed exactly as those of the initial problem solving test and included two new sequences and two sequences that were used in the problem-solving test."

In a study by Anastasiow _et al_. (1970), the retention test involved transfer items.

> The major test was employed as the pre-test, post-test and retention test. This 50 item instrument has five basic types of questions, namely matching by color and/or shape, recognizing and identifying the color and/or name of forms, completing sequences of forms, responding to void space and intersections, and verbalizing matching and intersection concepts. Included within the test are several transfer items. Most of the questions used the same three colors and three shapes employed in the curriculum which were called content items. The transfer items involved two sets of stimuli, either forms of three colors (brown, pink, or green) and three shapes (oval, rectangle, or diamond) or plastic utensils of three colors (pink, blue, or yellow) and three types (spoon, fork, and knife). None of these transfer items was used in the curriculum. (p. 498)

Concerning the relationship between transfer and retention, Gagne (1970) states:

> Learned capabilities that can be transferred must, of course, be stored, but they are not "recalled" in the sense that items of verbal information are. Instead, they are applied or used in a new situation. One may suppose, therefore, that the learner may need strategies for learning transfer that are different from those he uses in verbal recall. . . . (p. 77)

The use of the term retention varies from study to study. Rarely, in the research studies reviewed by this author, was the term defined; instead, its meaning was implicitly assumed.

### Transfer

Transfer is used in many different contexts. Bruner (1960), in the following passage, points out two types of transfer.

> The first object of any set of learning, over and beyond the pleasure it may serve, is that it should serve us in the future. . . . There are two ways in which learning serves the future. One through its specific applicability to tasks that are highly similar to those we originally learned to perform. Psychologists refer to this phenomenon as specific transfer of learning. . . . A second way in which earlier learning renders later performance more efficient is through what is conventionally called non-specific transfer or, more accurately, the transfer of principles and attitudes. In essence, it consists of learning initially not a skill but a general idea, which can then be used as a basis for recognizing subsequent problems as a special case of the idea originally mastered. This type of transfer is at the heart of the educational process. . . . (p. 17)

The first type of transfer that Bruner refers to is sometimes known as vertical transfer. Gagne (1970, p. 337) states that vertical transfer " . . . is observed when a capability to be learned is acquired more rapidly when it has been preceded by previous learning of subordinate capabilities." Bruner's second type of transfer is usually referred to as horizontal or lateral transfer. For Gagne (1970, p. 335), lateral transfer refers to " . . . a kind of generalizing that spreads over a broad set of situations at roughly the same level of complexity." For Hanson (1967, p. 24), transfer of learning " . . . refers to the applications of knowledge learned in one setting to problems in a setting which is only remotely similar to the setting in which the learning took place." According to Ray (1957, p. 5), "Transfer occurs when old learning and new problem situations are interrelated because of common components, factors, stimuli, or relations."

There are at least two forms of vagueness in transfer tests; variation in the different types of applications, and the degree of remoteness of these applications. Furthermore, in some studies it is a difficult matter to ascertain just what is being transferred in order to complete the learning task. For example, Bassler et al. (1971), in describing their horizontal and vertical transfer subtests, state:

> . . . The horizontal transfer subtest consisted of applications of the instructional materials to novel physical situations or to mathematical situations with a slight change in context. The vertical transfer subtest consisted of new and higher level mathematical tasks. In some cases a minimal amount of additional instruction was provided. This occurred when the students were told the meaning of an open sentence and its solution in the fourth grade. In other cases, it was a generalization of the instructional content such as the extension of the concepts

> of two dimensional vectors to three dimensional vectors in the eighth grade. (p. 307)

From this description, one has little more than a vague idea of the nature of the transfer tests.

In some situations, a learning test required the generalizing or specializing of learned material. Such is the case in the studies by Michael (1949) and Wolfe (1963). Michael compared an expository method and an inductive-discovery method of teaching the concepts and operations with signed numbers to ninth-grade algebra students. A generalization test was constructed to evaluate the following areas:

> 1. substitution of equivalents in general number expressions
> 2. interpretations of + and - numbers and zero in general form
> 3. using signed numbers in writing generalized expressions of oppositeness
> 4. determining the sign of the result of a process with general number expressions
> 5. writing a general expression of number relationships
> 6. interpreting a functional relationship in general number expressions
> 7. describing next steps in solution of equations (p. 86)

It is not clear how remote these applications are. Furthermore, one has difficulty in differentiating this test, as described, from an achievement test at the applications level. The remoteness of Wolfe's (1963) applications is clear from his transfer-test description.

> A transfer test was constructed to measure the ability of subjects to apply their knowledge of subject matter in situations which were unlike those in the programs. Some items required extrapolation from the concepts of generalizations developed in the instructional materials. For example, while the achievement test contained items dealing with the union of two sets as developed in the programs, the transfer test contained items requiring an extension of the concept of union to three or more sets. (p. 27)

In the study by Price (1967), an inductive reasoning test was employed. It is not clear whether this test should be considered as a transfer test. No description of the test was provided.

There are two degrees of transfer, positive and negative. Positive transfer involves the mastery of one task facilitating the mastery of another task. When a mastery of one task inhibits the mastery of another task, the transfer is called negative transfer. Concerning these two types of transfer, Shulman (1971) states:

> What is needed for positive transfer is to minimize all possible interference. In transfer of training, there are some ways in which the tasks transferred to are like the ones learned first, but in the other ways they are different. So transfer always involves striking a balance between these conflicting potentials for both

> positive and negative transfer. In discovery methods, learners may transfer more easily because they learn the immediate things less well. . . . (p. 189)

The majority of research studies reviewed for this study were concerned only with identifying positive-transfer powers. The study by Worthen (1968) was concerned with negative transfer.

Transfer of training is measured by many quantitative expressions. Because of their variety, the amount of transfer from one experimental study cannot necessarily be compared with the amount of transfer from another experimental study in any standard or systematic fashion. The majority of transfer studies reviewed by this author employed criterion tasks that were scored according to the frequency of correct answers or the amount of performance within a given time interval; these numerical values increase with learning. Other transfer measures were exemplified by such variables as the number of errors, time of response, and the number of trials to criterion; these scores decrease with improvement in performance. In a study by Hendrickson and Schroeder (1941), transfer was defined in terms of improvement from the first task to the second task; transfer was interpreted as the percentage of gain. In Swenson's (1949) study, transfer was measured by comparing each group's performance at different times during the experiment. Several criterion measures were used in the studies of Gagne and Brown (1961) and Meconi (1967). Both studies used three criterion measures to detect the effects of transfer: average time to criteria, the number of hints required in order to discover the rules, and a weighted time score.

In summary, a number of different meanings of transfer exist, and transfer is measured in a variety of different ways. As a result, it is difficult to compare transfer results between studies.

## Principles and Concepts

In a few studies there has been some confusion concerning the use of principles and concepts. A principle is usually thought of as a rule, generalization, or a prescription, and is built up from concepts. Concepts, on the other hand, according to Henderson (1967) and others, can be thought of as an ordered pair, the first component being a label and the second component consisting of a collection of meanings associated with the label. Of course, other conceptualizations of a concept exist. According to Hunt, Marin, and Stone (1966, p. 10), "A concept is a decision rule which, when applied to the description of an object, specifies whether or not a name can be applied." Thus, a concept is a principle under this definition. Following this convention, Hermann (1971, p. 24) states, " . . . The concept used was 'matriculation to Sydney University', which can be considered as a classification rule."

Keurst and Martin (1968), confusing concepts and principles, stated:

> . . . The problem for both groups was to learn the concept of finding the sum of a series of equally-spaced, consecutive numbers by the convenient technique of multiplying the middle number in the series by the number of members in the series. (p. 42)

Anastasiow et al. (1970) compared discovery and expository methodology for teaching mathematics to kindergarten students. They explained their experimental hypotheses by stating:

> It was predicted that 5-6 year-olds would, under guided discovery conditions learn the principles of set, intersection, form, and color with fewer errors and more correct verbalizations of the principles than would students taught under discovery and didactic teaching conditions. (pp. 494-5)

In this author's view, set, intersection, form, and color are concepts, not principles.

Hanson (1967) compared two programed methods (inductive-discovery and expository) of teaching certain content dealing with arithmetic sequences. In describing this content, he stated:

> . . . The following concepts are included in the study: arithmetic sequence; common difference and term of an arithmetic sequence; the use of a series of dots to represent missing terms in a long arithmetic sequence; arithmetic means; the determination of arithmetic means of two numbers; the writing of a term of an arithmetic sequence in terms of the first term, the position of the terms in the sequence, and the common difference; the determination of the $n^{th}$ term of an arithmetic sequence given the first term and the common difference; and the summing of the terms of an arithmetic sequence. (pp. 21-22)

The determination of arithmetic means of two numbers, the writing of a term of an arithmetic sequence in terms of the first term, the determination of the $n^{th}$ term of an arithmetic sequence given the first term and the common difference, and the summing of the terms of an arithmetic sequence involve rules or generalization and are not considered as concepts by this author. Each of these principles involve several component concepts. In order to verify that a student understands one of these principles, he is involved in demonstrating the principle or providing valid instances of the principle. To verify that a student understands a concept involves the student in classifying examples of the concept.

Worthen (1968, p. 225) used the term concepts rather loosely when he stated the following:

> The concepts selected were (1) notation, addition, and multiplication of integers (positive, negative, and zero), (2) the distributive principle of multiplication over addition, and (3) exponential notation and multiplication and division of numbers expressed in exponential notation. (p. 225)

There are several unfortunate consequences of the above semantic confusion. First of all, such confusion makes it more difficult to summarize the research findings. Secondly, a person inexperienced in dealing with such inconsistencies may formulate false concepts or impressions. Thirdly, such impreciseness does little to advance the art of communicating research findings to classroom teachers or other researchers. Last, but not least important, is the confusion that could result when teaching principles and concepts. Concepts have an arbitrary (definitional) nature, whereas the nature of principles is nonarbitrary. This difference makes a difference in teaching.

## Inductive Methods

Inductive teaching methods are often confused with discovery methods. But inductive strategies can be either discovery or expository. An inductive-discovery strategy proceeds from specific instances or examples to teach the object of the lesson. The student is not told the lesson objective, but must discover it. Questions are frequently asked by the teacher, and feedback may be provided.

An inductive-expository strategy differs from the inductive-discovery strategy in that the object of the lesson is stated by the teacher after the presentation of specific items of knowledge, e.g., in the case of teaching a principle, the teacher generalizes to the principle after presenting a sequence of instances of the principle. Hence, the main difference between the two methods is the teacher's statement of the result; in the inductive-discovery method, the result is not stated, at least at first or until verbalized by a student, and in the case of the inductive-expository method, the result is stated. In both cases further practice may follow.

The studies by Yabroff (1963), Denmark (1964), Eldredge (1965), Neuhouser (1964), Sheldon (1965), Lackner (1968), Stock (1971), and Sakmyster (1972) all involve inductive-expository strategies and have been cited as discovery studies. The study by Neuhouser (1964) involved three programed treatments, two "discovery" and one expository. In one discovery treatment, the students were required to verbalize their discoveries, while in the other, the students were not required to verbalize their findings. The nonverbalized discovery method was actually inductive-expository by the above criterion. Feedback, following each frame, was provided to the learners. All the above studies, with the exception of Stock's (1971), examined instruction in programed format which provided feedback to the learners, including statements of the generalizations. In these studies, there is no way of knowing whether the student discovered the generalizations or whether he learned them by reception.

## Deductive Methods

Deductive teaching methods are most often identified with expository methods. Teachers that employ the Socratic questioning method to get students to educe a proposition by juxtapositioning or combining known results are employing a deductive-discovery strategy. Such is the case with the studies by Levine (1967), Kuhfittig (1972), and Olander and Robertson (1973). Discovery, as with exposition, can be taught by deductive techniques as well as inductive ones.

## Methodological Problems in Experimental Designs

The comparative research literature on discovery has revealed a number of methodological problems in experimental designs. The more salient problems include the following: failure to *a priori* identify significance levels, failure to report power levels, using non-random selection procedures, failure to adequately control the teacher variable, failure to gather or report reliability information concerning the measuring instruments failure to control for pretest sensitization effects, and the use of short-term treatments. These problems will be discussed in the following sections.

## Post hoc Significance Levels

One generally uses a significance test when information about a population parameter is to be inferred from a single sample. It is ordinarily assumed that a sample (S) drawn from a population (P) will inherit certain characteristics from the population. The determinative influence of the population on the sample may be phrased as an implication "P implies S". It is the contrapositive of this implication, "not S implies not P", that provides the context for significance testing. Assuming that P possesses a certain parameter (null hypothesis), one tries to establish this assumption by drawing a random sample, computing a statistic which estimates the parameter, determining the frequency of occurrence of the statistic in the sampling distribution (since no two samples necessarily produce the same estimate), and then using a decision rule, which was formulated prior to gathering the sample, to determine the likelihood of the statistic's occurring. If the sample statistic is unlikely to occur, as determined by the decision rule, then, by the logic of contraposition, P does not possess the parameter; therefore the null hypothesis is rejected. The decision rule is stated as a particular significance level. Of course, this significance level is the probability of rejecting the null hypothesis when true. This whole inference scheme is based on only one sample being drawn.

Quite often it is the case that the p-value (the probability of obtaining a sample value at least as devious as the one obtained) is taken as the measure of significance. Concerning this, Bakan (1966) stated:

> A common misinterpretation of the test of significance is to regard it as a "measure" of significance. It is interpreted as the answer to the question "How significant is it?" A *p* value of .05 is thought of as less significant than a *p* value of .01, and so on. The characteristic practice on the part of psychologists is to compute, say, a *t*, and then "look up" the significance in the table, taking the *p* value as a *function* of *t*, and thereby a "measure" of significance. . . . But it must be remembered that this is using the *p* value as a *statistic descriptive of the sample alone*, and does not automatically give an inference to the population. (p. 428)

The p-value is a function of the type of test used and the sample statistic; if a new sample is drawn, a new p-value results. The significance level, on the other hand, is not dependent on the sample statistic or the type of test used.

The level of significance is not a measure that the population has a specified parameter; it either does, with probability one, or it doesn't, with probability zero. The level of significance is a decision rule, stated _a priori_, within the inference model. A significance level of .05 should be taken to mean that the chance of rejecting the null hypothesis when true is .05, whereas a p-value of .05 is an indication of the likelihood of drawing a particular sample.

A number of studies have either implicitly assumed the significance level or have confused it with the p-value, a poor procedure to follow in either case. Included among these are the studies by Hendrix (1947), Michael (1949), Kersh (1958), Gagne and Brown (1961), Foord (1964), Ballew (1967), Meconi (1967), Price (1967), Werdelin (1968), and Bassler _et al_. (1971).

The p-value should not be considered as a measure of significance or stated _a posteriori_. If the p-value is taken as the significance level, one is not adhering to the generally accepted inference model for making decisions concerning an experimental hypothesis.

## Power of Statistical Tests

There are two kinds of errors that can be made when testing a hypothesis: rejecting the null hypothesis when true (type I error) or accepting the null hypothesis when false (type II error).

The *power of a statistical test* is the probability that the null hypothesis will be rejected when false. Of course, if all hypotheses were rejected, then one would not need to be concerned with statistical power; but, unfortunately, this is not the case. Assuming a specific effect size for a fixed alpha level and a fixed sample size, a nonsignificant finding with power of .20 implies that the probability of accepting the hypothesis when false is .80. In effect, this adds support for the experimenter's conjecture.

Brewer (1972), commenting on the use of statistical power, stated:

> The only way for a researcher to be comfortable with a result is for his sample ES (expected size) and $\alpha$ to yield a large power. Then he knows that there is not much chance for an error of either kind and a good chance for a valid rejection. (p. 401)

The power of a statistical test is a very important concept, and has been nonexistent in the majority of the experimental research studies on discovery techniques reviewed by this author. Only the study by Caruso (1966) reported power levels for significance tests.

This author, in a preliminary study, made a review of the comparative research done on discovery teaching in mathematics from 1913 to the present. It was found that discovery teaching strategies were superior to expository strategies as judged by the number of statistically significant findings reported for both retention and transfer measures. For retention, approximately 50 per cent of the studies reviewed yielded no significant results, and for transfer, approximately 40 per cent of the studies reviewed yielded no significant results. For the transfer measures, every significant finding favored the discovery method. Since the majority of the studies did not report any power information, one cannot really interpret the nonsignificant findings; some may have resulted from tests with low power. Furthermore, one cannot determine the true state of affairs concerning a superior method of teaching. Either the discovery methods are superior, or the two general methods are equally effective, and the significant results were due to uncontrollable, confounding factors.

The reporting of power information is necessary for a well designed experimental study. The experimenter must attempt to control both types of errors.

## Nonrandom Sampling Procedures

Using nonrandom sampling procedures greatly ~~limits~~ the external validity of a research design. One of the major purposes of experimental research is to draw valid inferences concerning some population from a sample within that population. Kempthorne (1961) has distinguished between two types of populations: (1) experimentally accessible, and (2) target. The experimentally accessible population is the population of available students, and the target population is the totality of subjects that are of interest to the experimenter.

There are two degrees of external validity in any experiment: generalizing from a sample to the experimentally accessible population, and generalizing from the experimentally accessible population to the target population. Generalizing to the target population is difficult in practice. Concerning this, Bracht and Glass (1968, p. 441) state, "The degree of confidence with which an experimenter can generalize to the target population is never known because the experimenter is never able to sample randomly from the true target population." Valid inferences concerning the experimentally accessible population may be drawn from a sample provided the sample was randomly drawn from the population. Concerning nonrandom-selection procedures, Bracht and Glass (1968) state:

> If the sample has not been randomly selected from some experimentally accessible population, the experimenter cannot generalize with probabilistic rigor to some larger group of students. In reality, his sample has become his experimentally accessible population. (p. 441)

Using intact classrooms in experimental methodology studies is a questionable procedure. It is seldom the case that an intact class was formed without some selection process. The use of such classes results in experimental biases and leads to possible violations of assumptions requiring randomized procedures for certain statistical tests. If intact classes are employed, the experimenter should justify that the experimental procedure is free of dependence-producing effects, and also provide a rationale that the sample can be considered random; usually this results in "unnecessary" significance tests which tend to inflate the experimentwise error rate. Concerning sampling from intact groups, Peckham, Glass, and Hopkins (1969) state:

> Researchers would be well advised to achieve random assignment wherever possible, even with handicapped groups, where the variability is such as to make

> the investigator timorous of "randomization." The results of experiments involving intact groups should be regarded as thin evidence that a hypothesis deserves a more rigorous test. This is particularly true of experiments which employ only one intact group in each experiment. Increasing the number of groups may result in increased credibility of the basic hypothesis under study, but it should be recognized that using the individual as the unit of analysis when intact classes have been assigned to treatments is not logically or methodologically justifiable. (p. 338)

A number of discovery research studies used intact groups in their design. Among these are the studies of McConnell (1934), Thiele (1938), Michaels (1949), Sobel (1954), Fullerton (1955), Norman (1955), Caruso (1966), Nichols (1971), and Keese (1972). Hence, the results of these experiments should be considered with caution.

## Hawthorne and Novelty Effects

An examination of the experimental research on discovery has revealed many studies where the students were told that they were participating in an experiment. Among these are the studies by Ray (1957), Grote (1960), Moss (1960), Neuhouser (1964), Rowlett (1960, 1964), Ashton (1962), and Tomlinson (1962). The student's awareness that he is participating in an experiment may alter his performance. This phenomenon is referred to as the "Hawthorne" effect.

Bracht and Glass (1968) state several reasons why the Hawthorne effect might possibly contaminate experimental treatments. Certain students have high levels of "evaluation apprehension." Some are motivated by holding a high regard for the aims of science and experimentation or by just wanting to do the "right thing." There is some evidence (Cook, 1967) to suggest that the Hawthorne effect probably does not contaminate experimental results in measures of academic achievement to the extent claimed.

When students are subjected to a new or unusual experimental treatment, the observed behavior may be due to the treatment, the novelty factor, or both. According to Bracht and Glass (1968), an unfamiliar treatment may also have a disruptive effect.

> The antithesis of the novelty effect is the disruption effect which sometimes occurs with a new and unfamiliar treatment which is sufficiently different to the experimenter to render it somewhat less than effective during the initial try-out. After the experimenter has attained facility with the treatment, the results may be equal or superior to a traditional treatment. There is also the possibility that the novelty and disruptive effects counterbalance each other in the same experiment. (p. 459)

In the case of most discovery studies, the disruptive effect may be operating. Students who have long been exposed to expository techniques are suddenly exposed to a discovery strategy. The disruptive and novelty effects may become diluted or disappear completely over long periods of time. Unfortunately, most discovery studies have been short-term.

The disruptive or novelty effects may also be operating when different strategies are presented in programed form, particularly if the students have not been exposed or accustomed to programed instruction previously.

## Control of the Teacher Variable

The teacher variable has not been adequately controlled in many discovery experiments where the instruction has not been in programed form. In some studies (e.g., Caruso, 1966; Fullerton, 1955; and Keese, 1972) an attempt is made to control the teacher variable by having each involved teacher instruct both an expository and a discovery group. But with this method, if the teachers are not adequately familiar with discovery teaching techniques, or if they are not randomly assigned to the treatment classes, confounding or contamination of the treatments may result. The average teacher is generally not familiar with discovery teaching techniques.

Other studies (e.g., Craik, 1966; Levine, 1967; Babikian, 1971; Cooke, 1971; and Kuhfittig, 1972) attempted to control the teacher variable by having the experimenter teach all the treatment groups. In these cases, experimental bias may have resulted from the "Rosenthal" effect. Kuhfittig (1972, p. 43) uses the term Rosenthal effect to describe the effects that may result when " . . . the teacher may unintentionally bias his instruction in the direction of confirming the hypothesis."

The studies by Ray (1957), Moss (1960), Grote (1960), and Rowlett (1960, 1964) attempted to partially control the teacher variable by using tape recorded instruction, creating an "artificial" atmosphere for the period of instruction.

The study by Hirsch (1972) employed several schools with two discovery classes at each of two schools. At each of these schools, the instructor of the class observed the investigator teach the first class and then taught the second class using the pedagogical procedures of the investigator. It was not reported whether the investigator observed the subsequent teaching behavior of the instructors. The novelty effects of a new teacher may have resulted in experimental bias. This could not be determined from the data reported since the data resulting within each school were pooled to form one group of scores.

The study by Kuhfittig (1972) used audio tapes to insure uniformity of instructional procedures. With audio recordings, one could not assess the nonverbal behavior of the teacher.

Rizzuto (1972), in a seven-week study, video taped each of his teachers on two separate occasions to insure teacher fidelity to treatments. Since the teachers undoubtedly knew they were being taped, they were probably at their best behavior. It is questionable how this small sample of behavior could have been a valid estimate of behavior for the total experiment.

## Reliability and Validity Data

The reliability of a measuring instrument has a direct effect on the validity of the inferences drawn from the data. Yet much experimental research literature on discovery reports no validity or reliability data concerning their measuring instruments. For example, the studies by Kersh (1958), Foord (1964), Craig (1965), Price (1965), Krumboltz and Yabroff (1965), Belcastro (1966), Ballew (1967), Hanson (1967), Keurst and Martin (1968), Jamieson (1969, 1970, 1971), Anastasiow *et al*. (1970), Barrish (1970), Bassler *et al*. (1971), and Cooke (1971) fail to report any reliability or validity information. This is not an exhaustive list, but used to indicate that a problem does exist.

The measuring instrument could be considered the most important part of the experimental design. With a weak criterion measure, one could be measuring some other behavior or measuring the behavior in an inaccurate fashion. In either case, the results could be misleading. Thus, if valid inferences are to be drawn from the data, the data must result from valid and reliable measures.

A number of studies (e.g., Swenson, 1949; Ashton, 1962; and Nichols, 1971) employed gain or difference scores to analyze their data. It is well known that difference scores are generally less reliable than either of their components. Concerning this, Thorndike and Hagen (1961) state:

> It is, unfortunately, true that the appraisal of the difference between two tests usually has substantially lower reliability than the reliability of the two tests taken separately. This is due to two factors: (1) the errors of measurement in both separate tests effect the difference score, and (2) whatever is common to both measures is canceled out in the difference score. (p. 191)

Brown (1970) cautions the use of difference scores.

> . . . Extreme caution must thus be used in interpreting difference scores, especially when, as is often the case, an index of the reliability of the difference is not readily available. (p. 89)

The study by Ashton (1962) employed a measure which had an average (resulting from ten groups) internal-consistency reliability estimate of .50, ranging from .02 to .81. As a pretest, the average reliability estimate was .55, ranging from .22 to .79; and as a posttest, the average reliability estimate was .62, ranging from .02 to .81. With these reliability estimates, it is difficult to interpret her data involving pretest-to-posttest gain scores.

Pretest Sensitization Effects

Pretesting may limit the generalizability of an experiment and confound the treatment effect. Students exposed to a pretest may be no longer representatives of the accessible experimental population if pretested unless the design controls for pretest effects. Concerning the effects of pretest sensitization, Bracht and Glass (1968) stated:

> The results of empirical investigations of pretest sensitization indicate that the effect is most likely to occur when the dependent variable is a self-report measure of some aspect of personality, attitude, or opinion. The pretest effect on academic achievement is apparently less prevalent, but the results are inconclusive since the studies which have been conducted are not representative of experimental situations where it usually is necessary to use a pretest. (p. 463)

To this author's knowledge, of the many experimental studies comparing discovery and expository methods of instruction and employing pretests, only the study by Kanes (1971) considered the possible pretest effects; all students were not exposed to the pretest. The results indicated that there were no significant differences between the groups pretested and those not pretested as measured by a posttest.

A number of research studies involved pretests, but did not report controlling for pretest sensitization effects.

Among these are the studies by Michael (1949), Fullerton (1955), Norman (1955), Ballew (1965), Shelton (1965), Craik (1966), Lennek (1967), Brenner (1968), Lackner (1968), Strickland (1968), and Babikian (1971).

## Length of Studies

A large number of experimental research studies investigating discovery strategies have been short-term studies, the instructional period being less than three hours. In fact, many studies (Ray, 1957; Grote, 1960; Moss, 1960; Rowlett, 1960, 1964; Tomlinson, 1962; Kufittig, 1972; and others) employed a longer examination period than instructional period.

Concerning the briefness of instruction in discovery studies, Cronbach (1966) stated:

> . . . Studies of inductive teaching have generally employed very brief instruction, yet the recommendations apply to whole courses or whole curricula. . . . Even as small experiments, the discovery studies have been too miniature. Typically, there is an hour of training and one delayed transfer test. (p. 86)

By employing short instructional periods, the novelty or disruptive effects do not have a chance to diminish.

There is some research evidence to support the claim that short-term studies in transfer may lead to misleading conclusions. Duncan (1964) conducted a learning study with

100 college students using two kinds of 10-item verbal lists. The stimuli consisted of the digits 1-10 and were paired with adjectives. Half of the students learned paired-associate lists containing one response per stimulus. The remaining students learned response-discovery lists where each stimulus could be paired with three possible responses, only one of which was correct per stimulus. The learner had to guess which adjective had been selected randomly as correct. Discovery was by trial and error. Each student learned five lists, one list per day for five days. Duncan found that with students studying the paired-associate lists, there was negative transfer (comparing mean correct responses per trial) from day 1 to day 2, but an overall gain (learner to learner) from day 1 to day 5. Thus, if this study would have been confined to two days, misleading or false conclusions would have been drawn.

## Problems in Data Analysis

A review of at least 60 comparative research studies investigating treatment effects on the ability to retain or transfer learned knowledge has indicated that a number of different procedures have been used to analyze the data. For examining main effects, the majority of studies employed either a randomized ANOVA model (e.g., Fullerton, 1955;

Yabroff, 1963; Dennison, 1969; and Sakmyster, 1972) or a t-test model (e.g., Hendrix, 1949; Ashton, 1962; Neuhouser, 1964; and Keurst and Martin, 1968). ANCOVA models (e.g., Eldredge, 1965; Scott, 1970; and Olander and Robertson, 1973) and repeated measure designs (e.g., Kuhfittig, 1972; Murdoch, 1971; and Cooke, 1971) were also used but with less frequency.

Pretests were sometimes used to block the data into levels (e.g., Shelton, 1965 and Lackner, 1968) or to determine aptitudes or background knowledge (e.g., Sobel, 1954 and Brenner, 1968). Occasionally, the pretest score was used as a covariate in an ANCOVA design (e.g., Shelton, 1968; Michael, 1949; Lackner, 1968; and Olander and Robertson, 1973).

Control groups were included in some designs to control for maturation factors and to provide baseline information for making comparisons or determining the difficulty of criterion measures (e.g., Schaaf, 1954; Ray, 1957; Moss, 1960; and Rowlett, 1960).

A number of controversial procedures and questionable methodology for analyzing data were detected. Among these, the most glaring were (1) choice of experimental unit, (2) violations of the ANOVA model when $F < 1$, (3) covariates affected by the treatment in ANCOVA designs, (4) indiscriminate pooling of data, (5) improper post hoc comparisons, (6) multiple comparisons and questionable error rates, (7) use of inappropriate statistical models, and (8) pre-study investigations and homogeneity of variance assumptions. These problems will be discussed in the following sections.

## Experimental Unit

Both the t-test for uncorrelated samples and the randomized ANOVA model for single factor experiments satisfy the simple, linear structural model $X_{ij} = \mu + \alpha_j + e_{ij}$ where $X_{ij}$ represents the element i in sample j, $\alpha_j$ is the effect of treatment, j, $e_{ij}$ is the experimental error, and $\mu$ is the grand mean of the treatment populations. For each j, the $e_{ij}$ must satisfy the following properties: (1) $\alpha_j$ and $e_{ij}$ are independent, (2) the $e_{ij}$ are independent, (3) $e_{ij}$ is a normally distributed random variable, (4) $e_{ij}$ has a common error variance for the J populations (for the t-test, $J = 2$), and $E(e_{ij}) = 0$. The experimental error, $e_{ij}$, represents all uncontrolled sources of variance affecting the measurement of $X_{ij}$. $\mu$ is a constant, and $\alpha_j$ is constant for all elements in population j. The ANOVA model is robust with respect to minor violations of the assumptions of homogeneous variances and normality when equal sample sizes are involved (Glass and Stanley; 1970, pp. 371-72). The effects of violating the independence of the $e_{ij}$ assumption are controversial.

The experimental unit is usually regarded as the smallest subgroup within an experiment that may receive different treatments. A more concise definition is given by Peckham _et_ _al_. (1969).

> . . . The experimental units are the smallest divisions of the collection of the experimental subjects which have been randomly assigned to the different conditions in the experiment and which have responded independently of each other for the duration of the experiment, or which, if allowed to interact during the experimental period, have had the influence of all extraneous variables controlled through randomization. (p. 341)

A strict interpretation of this definition would almost invariably result in the class mean becoming the unit of analysis.

The task of choosing the experimental unit is not always an easy one; determining the validity of the independence of scores or replications requires considerable judgment on the part of the experimenter. If one adopts the class mean as the experimental unit, he is deprived of stratifying on certain variables and cannot check for significant interaction effects of factors. A use of the class mean generally results in reduced power to detect differences. On the other hand, Peckham _et_ _al_. (p. 344) indicate that the variation among class means is much less than among individual pupils within classes. The reduced power is reflected by the loss in degrees of freedom for the error term involved in the F-test. To some extent, this is compensated for by a decrease in the size of the error term, resulting in a larger F-value.

There is evidence to suggest that violating the independence-of-errors assumption may substantially affect the validity of probability statements; in most cases, the probability of committing a type I error is decreased and results in a loss of power. A study by Worthen (1968) used the pupil as the unit of analysis and achieved significance on a number of criterion measures. When his analyses were criticized and reanalyses were carried out using the class mean as the unit of analysis, there were no treatment differences on any of the dependent variables (Worthen and Collins, 1971).

Steck (1966) conducted a study comparing the scores made on an achievement test by two groups of students, one group receiving instruction on a one-to-one basis while the other group received instruction under a group format. To insure uniformity of treatment, tape recorded instruction was used. Steck found that the students who learned in a group had a significantly smaller score variance than those learning individually. It stands to reason that students become less variable when exposed to the same experiences. Concerning this phenomenon, Raths (1967) stated:

> If the variability of the group is thus diminished by common experiences, then statistical tests of the mean differences are inappropriate since these tests assume independence of the data. Unfortunately, the violation of the independence function works to yield more significant results than should be expected by chance, so researchers who use individuals and not classrooms as the units in methodology studies generally report spurious significant results associated with their analysis. (This problem of appropriate units is not encountered by researchers who use treatments which present material to students to students one at a time--such as programed learning). (p. 265)

A number of experimental research studies (e.g., Boeck, 1951; Ashton, 1962; Naughton, 1962; Caruso, 1966; Worthen, 1968; Barrish, 1970; Rizzuto, 1970; and Babikian, 1971) have used questionable experimental units in analyzing their data.

### Violating the ANOVA Model when $F < 1$

Considering the ANOVA model, if the null hypothesis of equality of treatment population means is true, then $E(MS_b) = E(MS_w)$ and $E(F) = 1$. If treatment effects exist, $E(MS_b)$ is greater than $E(MS_w)$ and $E(F)$ is greater than one. Only an F-statistic greater than one will offer evidence that the null hypothesis is false; of course, this difference may have occurred by chance or as a result of sampling errors. If the observed F-value exceeds the critical F-value, then it is concluded that the observed differences are not likely to be due to chance factors alone, and the null hypothesis is rejected; treatment differences do exist.

Generally, if an F-statistic is less than one, the F-test is considered nonsignificant. The F-test that is employed in the ANOVA designs is a one-tailed test. If F is less than one, it may be the case that the F-statistic is significantly small, i.e., F is less than the critical value for the lower 5 per cent tail. In this case, one cannot reject the null hypothesis, but must assume that the ANOVA model has been violated in some way. Concerning this, Myers (1966) states:

> . . . The occurrence of F's so small that their reciprocals are significant or the occurrence of many F's less than one in a single analysis of variance merits further consideration. Such findings suggest that the model underlying the analysis of variance has in some way been violated. (pp. 66-67)

Evidently, when this occurs, the treatments contain some systematic factor that makes the groups more homogeneous.

A number of studies (e.g., Sobel, 1954; Fullerton, 1955; Moss, 1960; Rowlett, 1960; and Lahnston , 1972) were found that reported F's that violeted the ANOVA model, according to Myers' criteria.

## ANCOVA and Confounding Covariates

There are two general methods for controlling the differences existing between experimental groups, experimental and statistical. The experimental methods precede the treatment and involve homogeneous blocking or stratifying of experimental units across treatments; in effect, reducing the variability within treatment groups and increasing the accuracy of the criterion. The statistical methods control for undue variability by removing sources of contamination of the dependent variables after the treatment has been administered. The analysis of covariance (ANCOVA) is a blending of regression analysis and analysis of variance (ANOVA), which provides statistical controls rather than experimental. When concomitant variables, such as pretest scores, aptitude scores, or I.Q. scores, can be identified that are correlated with the criterion variables, the portion of the variance in criterion scores caused by their effects can be partialed out by using ANCOVA, resulting in a score variance caused mainly by the treatments.

The ANCOVA model is based on the assumption that the additive components (e.g., treatment effect, covariate component, and error component) are statistically independent (Evans and Anastasio, 1968; p. 231). Problems exist when the covariates are influenced or correlated with the treatment; when the covariate is partialed from the analysis, spurious significant results may be obtained. Concerning this, Winer (1962) states:

> When the covariate is actually affected by the treatment, the adjustment process removes more than an error component from the criterion; it also removes part of the treatment effect. (p. 580)

Evans and Anastasio (1968) present an example where the treatment and covariate are correlated and result in the covariate variance being transmitted to the variate; the ANOVA resulted in a nonsignificant test, whereas the ANCOVA resulted in a significant test. Further, they state that a usage

> . . . in which the treatment effect is correlated with the covariate, is an inappropriate application of the ANCOVA because it violates a basic assumption of the model, that of independence of additive components. Consequently, homogeneity of within-group and between-group regressions is precluded and the treatment means may be overadjusted or underadjusted. In either case, the results would be spurious. (p. 233)

A study by Worthen (1968) used ANCOVA with a questionable covariate; the Concept Knowledge posttest was used as a covariate in the analysis of the other posttests and retention tests. A number of significant differences were reported. When the questionable covariate was deleted and a reanalysis conducted, all significant differences between treatments vanished (Worthen and Collins, 1971).

Some experimenters (e.g., Michael, 1949) used ANCOVA with intact, nonrandomized classes. In such cases, if the covariate is unreliable, there is evidence to suggest that spurious results can be generated (Cochran; 1968, p. 653).

A number of research studies were examined that employed the ANCOVA model to analyze their data where it seemed to this author that the treatment and covariate were related. Included among these studies are the studies by Eldredge (1965), Sheldon (1965), Lackner (1968), Worthen (1968), Brown (1969), and Scott (1970). In the studies by Shelton (1965) and Lackner (1968), the covariate (pretest scores) was used to define the two levels of ability used in each study. For each study, since the treatments differentially affected ability levels which were defined in terms of the pretest scores (the covariate), obscured results were possible after the covariate was partialed out.

## Indiscriminate Pooling of Data

When analyzing experimental data, many extraneous factors can intervene. Concerning this, Dixon and Massey (1969) state:

> In sampling from two populations it sometimes happens that extraneous factors cause a significant difference in means, even though there may be no differences in the effects we are trying to measure. Conversely, extraneous factors can mask or obscure a real difference. For example, in an experiment

> to test which of two types (A or B) of fertilizers is the better, two plots of wheat are planted at each of ten experimental stations. One of the two plots has fertilizer A and the other fertilizer B. If the average yield of the 10 having type A is compared with the average of the 10 having type B, part of the difference observed (if there is any) may be due to the different types of soil or different weather conditions at the different stations instead of the different fertilizer. Or the fertilizer may cause a difference, which is obscured by the other factors. (pp. 119-20)

The above paradigm and extentions, while not concerned with fertilizers, were encountered in methodology studies comparing discovery and expository teaching methods. Instead of comparing fertilizers on wheat, teaching methods were compared on students. An equal number of classes were usually assigned to each treatment, and after the treatments were administered, a posttest followed. The methods were then compared by combining class scores within each treatment and then comparing the grand means between treatments, a questionable procedure to follow as was pointed out above.

Two procedures could be followed to remove some of the doubt in the above procedures of pooling data. The first involves a demonstration that the means and variances within each treatment do not significantly differ. The other method involves matching classes across the two treatments.

Of the studies that involved pooling-of-data procedures, only the studies by Sobel (1954) and Anderson (1949) provided justification for pooling their data. In Anderson's study, he indicated that

> . . . it was necessary to pool the data collected in the several classes taught by a given method. Before the data could be pooled validly, it was necessary to determine whether the means and variances of the respective classes were homogeneous. (pp. 48-49)

Sobel also demonstrated that class means and variances within each treatment were homogeneous prior to pooling class data within each treatment.

The study by Babikian (1971) illustrates the most frequent application for pooling data that this author encountered in his review. Three teaching techniques were compared. The students were randomly assigned to nine classes of 24 students each; each class contained an equal number of boys and girls and students from two ability levels. One instructor taught all classes for a one-week period, three classes assigned to each method. One method was taught one week, one the next week, and the last the third week. The data were analyzed by using a 3x2x2, methods-by-levels-by-sex, ANOVA design. The design used to analyze the data assumed that 12 treatment groups were involved (18 students per cell), but only nine groups were involved in the experimental design. Furthermore, no justification was reported for this pooling of the data into the 12 cells. By pooling his data, many extraneous factors could have possibly been introduced which confounded his analysis. For example, history or time effects in one class may have produced extreme scores for certain students, and these scores, when pooled, could have significantly altered a main-effect mean.

A number of the studies investigated by this author employed questionable procedures of pooling data. Among these are the studies by Grote (1960), Moss (1960), Naughton (1962), Rowlett (1964), Caruso (1966), Barrish (1970), Babikian (1971), and Nichols (1971). As a result, their findings should be considered with caution.

### Improper Post Hoc Comparisons

Following a significant F-test for treatment differences between three or more groups, the problem becomes one of comparing each group with every other group and then assigning a level of significance to the conclusions. It is generally well known that multiple-t comparisons to check for between-group differences, either in place of an F-test or following a significant F-test, is an inappropriate procedure to follow and leads to a questionable probability model on which decisions are based. Concerning pairwise t-comparisons, Hays (1963) states:

> One is never really justified in carrying out the $\binom{J}{2}$ different t-tests for the differences among J groups, . . . such t-tests carried out on all pairs of means must necessarily extract redundant, overlapping, information from the data, and as a result a complicated pattern of dependency must exist among the tests. Furthermore, the apparent levels of significance found from a set of such tests have neither a simple interpretation nor a simple connection with the hypothesis tested by the F test. (p. 375)

Roscoe (1969, p. 239) states that the t-test is "inappropriate for testing the significance of the difference between any two means from a collection of three or more samples." If five different groups were involved in a significant F-test, there would be a total of 10 possible t-comparisons made at, say, the .05 level. If these t-tests were independent, which they are not, the probability of committing at least one type I error is $1 - (.95)^{10} = .40$. Since the t-tests are not independent, the true probability must lie somewhere between .05 (the level for one comparison) and .50 (by Boole's inequality).

The studies by Hendrix (1947), Grote (1960), Gagne and Brown (1961), Tomlinson (1962), Rowlett (1964), Neuhouser (1964), Ballew (1967), Britton (1969), and Richardson and Renner (1970) used pairwise t-comparisons instead of a more appropriate design (such as ANOVA) for detecting between group differences.

A number of studies (e.g., Swenson, 1949; Norman, 1955; Naughton, 1962; Luck, 1966; Babikian, 1971; and Gabor, 1972) followed a significant F-test with pairwise t-comparisons to detect pairwise differences.

## Multiple Comparisons and Error Rates

The concept of significance level has proved to be useful in dealing with experiments involving a single test, e.g., a t-test, a z-test, an F-test, or a chi-square test. When multiple comparisons become involved in a single experiment, such as several F-tests and several tests comparing the pairwise differences of means, the concept of significance level becomes obscured. Here, the problem is one of evaluating the type I error. Three types of error rates have been identified by Ryan (1959): error rate per comparison, error rate per experiment, and error rate per experimentwise. The *error rate per comparison* is defined as the probability that any given comparison will be declared significant when the null hypothesis is true. The *error rate per experiment* is the expected number of errors per experiment; this number could exceed one. The *experimentwise error rate* is defined for experiments containing multiple comparisons and is the probability that at least one comparison will be declared significant when the null hypothesis is true for all comparisons. For single experiments with a single comparison, the three error rates yield the same information; they become more divergent as the number of comparisons increases.

A great majority of the experimental research studies on discovery revealed upon review that no attempt was made to control the experimentwise error rate by choosing appropriate or adequate designs. In studies determining the effectiveness of several methods of instruction as determined by several criterion variables, significance tests were conducted to ascertain that groups were matched prior to the treatment, to ascertain that homogeneity assumptions were satisfied, to test for main effects on each criterion variable, and to locate pairwise differences after a significant F-ratio.

The study by Sobel (1954) is representative of those comparative research studies that contain questionable experimentwise error rates and a large number of significance tests. Sobel compared a deductive-expository method and an inductive-discovery method for teaching certain algebraic concepts and skills to 14 classes of ninth-grade students. Prior to the experimental treatment, an inspection of the mean I.Q. for each class indicated that each class could be identified as being average or high in intelligence. As a result, nine average-ability and five high-ability classes were identified. The two treatments and two ability levels resulted in four treatment combinations, each combination involving at least two classes.

A posttest, consisting of two parts, to measure concepts and skills was given at the conclusion of the experimental treatments. A parallel form of the posttest was given three months later as a retention test.

For each criterion measure, the data within each treatment combination was pooled after verifying that homogeneity of means and variances were satisfied at the .05 level of significance. A total of 40 significance tests resulted. Of these, two would be expected to be rejected by chance factors alone, if independent.

After pooling the data, the final analysis employed eight one-way ANOVA tests, four for the posttest and four for the retention test. For each criterion measure, assuming independence of significance tests, the probability of committing at least one type I error becomes .19 ($.19 = 1 - .95^4$) instead of the experimental standard of .05. For the eight significance tests, if independent, the probability of committing at least one type I error is .34. Of course, the significance tests are not all independent. Therefore, one cannot establish a precise error rate. For the eight significance tests, the probability of committing at least one type I error is somewhere between .05 (the alpha level) and .40 (by Boole's inequality).

Sobel's study contains at least 48 significance tests. It is impossible to identify the size of the experimentwise error rate. His experimental design, evidently, was not chosen to minimize this error rate. A more efficient design could have been formulated to test his hypotheses.

If p individual significance tests were independent of one another, the problem of determining the experimentwise error rate is rather straightforward; but, usually this is not the case. Whether or not the comparisons are independent does not affect the error rate per comparison (Ryan, 1959). When the significance tests are dependent, as is the case of an experiment where p dependent variables are correlated, the actual experimentwise error rate is usually unknown (Bock and Haggard, 1968). Hummel and Sligo (1971) conducted a Monte Carlo study to empirically study various error rates for certain numbers of criterion variables with varying degrees of dependence. Concerning the use of univariate tests, they state:

> The use of the univariate approach with multivariate data should be discouraged, particularly as p (the number of dependent variables) and proportion of variance in common increase. The grouping of errors and the generally unknown experimentwise error rates can easily allow for misinterpretations of the findings. (p. 56)

Thus, the many comparative research studies examined and containing at least two dependent variables stand a chance of being misinterpreted due to questionable experimentwise error rates.

## Inappropriate Models

In reviewing the research studies comparing the effectiveness of discovery and expository techniques of teaching, it was found that the majority involved at least two criterion variables. Concerning the effects on retention and transfer, a number of studies contained at least five dependent variables. Among these are the studies by Ray, 1957; Rowlett, 1960, 1964; Moss, 1960; and Grote, 1960. Not a single study (of approximately 100 reviewed) used a multivariate analysis of variance (MANOVA) design. Tatsuoka (1969) stated that the use of a series of univariate tests to evaluate multivariate data is not a valid procedure to follow.

> . . . for one thing, the statistical dependence among the several criterion variables upsets the significance levels in the series of univariate tests. (p. 740)

It is possible that a series of univariate tests would lead to the acceptance of each null hypothesis, and a multivariate test leading to rejection of a hypothesis of equal centroids. Also, it is possible that a series of univariate tests would each reject a null hypothesis and a multivariate test would accept a hypothesis of equal centroids of means. Concerning the appropriateness of the MANOVA model, Tatsuoka (1969) stated:

> Any time the experimental design is such that an analysis of variance (ANOVA) of some type (one-way classification, multi-factor design with crossed or nested factors, Latin-square design, etc.) would be appropriate if there were but one dependent variable, then MANOVA is applicable when there are two or more dependent variables. (p. 740)

### Pre-study Investigations and Homogeneity of Variances Assumptions

The t-test and the F-test are robust tests. Robust, a term invented by Box (1953), is a property of significance tests characterized by non-sensitivity to minor violations of normality and homogeneous variances assumptions, but is sensitive to the falsity of the null hypothesis it is used to test. Concerning the robustness of the F-test, Glass and Stanley (1969, p. 371) state, "When the sample sizes are equal, the effects of heterogeneous variances on the level of significance of the F-test is negligible."

When testing for homogeneity of variances, the larger the alpha level, the more conservative the test. For the majority of applications, the .05 level is chosen; a few studies use the .01 level. In testing the null hypothesis of equality of variances, the experimenter wants to accept the null hypothesis, not reject it, as is usually the case. The .05 level seems to this author to be too liberal a level to detect variations, particularly when unequal n's are involved. For such tests, one should be more concerned with the type II error (accepting the hypothesis when false) rather than the type I error (rejecting the hypothesis when true) since a type II error is probably the more serious error to commit in this case. Unequal groups could lead to spurious results in methodology studies. Concerning a statistical convention for the type II error, Cohen (1965) states:

> Statistical conventions, although frequently misused, are nevertheless useful, and I would suggest that if a conventional value for $\beta$ is desired, .20 be taken, i.e., the power of .80 be sought when no other basis is available. Like all conventions, this value is arbitrary, but it is, I believe, reasonable. (p. 98)

Further, when one wants to reject a hypothesis, and a more rigid test is desired, one chooses the .01 level rather than the .05 level. Similarly, if one wants to accept a hypothesis, and a more rigid test is desired, one should choose the .10 level rather than the .05 level. This stands to reason.

For the many studies that used pretests to determine whether groups were matched prior to being exposed to the treatments, the .05 level of significance was used (if stated). Here the experimenter also wanted to accept the hypothesis of equal means; the smaller the significance level, the better the experimenter's chances of accepting the null hypothesis. A conservative level such as .20 would make discovery of inequality of groups more likely, since inequality of groups could seriously affect the validity of the analysis used.

## Summary

This chapter indicates that the experimental research studies on discovery teaching and discovery learning contain a number of controversial or questionable procedures; the inferences drawn from these studies are the responsibility of the reader. Before a research finding is quoted, a thorough inspection of the research report should be made, including the data analysis, to determine the conditions and limitations of the study and whether the findings have any external validity. The study by Hendrix (1947) is often quoted as supporting the superiority of the subverbal awareness approach to student discovery as determined by a transfer measure. Yet this finding, as reported by Hendrix, was significant at the 12 per cent level, far in excess of the .05 experimental standard for developmental studies.

It is extremely difficult to adapt empirical data to a statistical research model. Therefore, if a finding has not been misrepresented by an experimenter and an invalid inference is drawn and reported by a reader of the research, the responsibility for any consequences of misrepresentation should be with the reader.

# CHAPTER IV

## REVIEW, ANALYSIS, AND SUMMARY OF THE RESEARCH

This chapter will be devoted to reviewing, analyzing, and summarizing the comparative (discovery versus expository) research studies investigating retention or transfer. The investigation of the research will be summarized separately for both mathematical and non-mathematical subject areas. The non-mathematical subject areas using school-related learning tasks are science, languages, industrial education, and geography. For both categories, a separate investigation will be conducted for retention and transfer at each of four levels. These levels are elementary (grades K - 6), junior high (grades 7 - 9), high school (grades 10 - 12), and college. The studies will be reviewed in chronological order.

Only those comparative studies that include at least one discovery teaching strategy and at least one expository teaching strategy will be included in this investigation. For the purposes of this study, a teaching strategy is a discovery strategy provided that the object of the lesson (a concept, a principle, etc.) is not explicitly told, either verbally or didacticly, to the learner. On the other hand, under an expository teaching strategy, the learner is told sometime during the lesson the objective of the lesson. Of course, in an expository strategy, the learner could discover the lesson objectives before being told by the teacher; but, in this case, one cannot always be sure of such discoveries or when they take place, if indeed they do. In any case, discovery learning is difficult to establish, and is not the main concern of this study; discovery teaching is the object of main consideration.

Comparative studies will be used to denote discovery versus expository studies.

## Mathematics

### Retention

Retention of learning is a two-stage process, storage and retrieval. In order to retrieve learning it must be stored. If certain learning is to be stored at time $t_1$, and a retention test is administered at time $t_2$, one should not assume that the learning was retained over the time interval $(t_1, t_2)$. It may be the case that the learning occurred somewhere between time $t_1$ and time $t_2$ instead of occurring at time $t_1$. Thus, only those comparative research studies that investigated initial learning (storage) and delayed recall were considered for investigation by this author. In the studies reviewed, the retention measure (delayed recall) was generally administered from two to six weeks following the initial achievement test.

Thirty-five comparative studies have been identified that investigated the retention of mathematical learning.

Elementary level (grades K - 6). Table 1 contains a list of the studies investigating retention of mathematical learning at the elementary level. Fourteen of these studies reported statistically significant findings concerning retention of mathematical learning. Those studies which yielded statistically significant findings concerning retention will be reviewed and analyzed in a chronological order. Unless a study yielding no significant results concerning retention has a serious defect in design or data analysis, it will not be reviewed under this section.

Winch (1913), in what appears to be the first experimental study concerning discovery teaching in mathematics, performed five experiments, all in different schools, using children ranging in age from eight to fifteen years to determine the relative effectiveness of inductive-discovery and deductive-expository methods of teaching certain

## Table 1

### Method Favored as Determined by the Significant Findings Reported by Experimenters for Retention of Mathematics

### Elementary Level

### (Kindergarten to Sixth Grade)

| Discovery | | Expository | Neither |
|---|---|---|---|
| Cooke (1971) | | Barrish (1970) | Anastasiow *et al.* (1970) |
| Fullerton (1955) | | Keurst and Martin (1968) | Bassler *et al.* (1971) |
| McConnell (1934) | | Murdoch (1971) | Kanes (1971) |
| Nichols (1971) | | Olander and Robertson (1973) (applications) | Kersh (1965) |
| Norman (1955) | | | Peters (1970) |
| Olander and Robertson (1973) (computations) | | Winch (1913) | Swenson (1947) |
| Scott (1970) | | | |
| Thiele (1938) | | | |
| Werdelin (1968) | | | |
| Worthen (1965) | | | |
| Total | 9.5 | 4.5 | 6 |
| Per cent* | 48 | 22 | 30 |

*Based on a total of 20 studies.

geometrical definitions. In the deductive method, definitions were given, usually written on the board, with illustrative examples following. The inductive method followed a form of Socratic questioning. With the definition in mind, the teacher, by using examples at the chalkboard, taught up to it; no instruction was given by the teacher other than by questioning. Winch stated:

> With children taught frequently on this method it is quite possible to get the necessary drawings and corrections, or most of them, done by members of the class, so that the machinery of correction and amplification is mainly in the hands of the class, with the teacher there to see fair play and direct the discussion to profitable issues. (p. 34)

The study included such definitions as square, triangle, oblong (rectangle), and diameter of a circle. Tests of immediate acquisition, retention, and transfer were given to measure the effectiveness of instruction. The transfer tests required the students to write definitions of new geometric concepts after having been exposed to written examples of the concepts. The tests were scored according to the number of correct and incorrect attributes identified for each definition. On immediate learning and two-week retention tests, Winch found that students exposed to deductive methods were superior, while on two-week transfer tests, the inductive groups were superior. The data were analyzed using difference scores and probable errors of these differences. A finding was deemed significant if the difference between two group means was greater than or equal to twice the probable error of the differences. The posttest and retention tests were repetitious of the pretests.

This study has many weaknesses. Among these are the use of non-random selection procedures, the lack of control of intervening variables, the use of criterion measures with questionable reliability, and the data analysis. Winch's study is a weak experimental study, judged by today's standards. This study was primarily included in this report for historical purposes.

In a study lasting more than seven months, McConnell (1934) taught 100 basic addition and 100 basic subtraction facts to more than 1,000 second grade students from 15 schools using two methods. The two methods were identified as the authoritative and discovery methods. In the authoritative method, the number combinations, considered as stimulus connections, were identified without meaning; the student took the word of the teacher that his work was wrong and depended on the teacher to supply him with the correct answer. The students were to learn the number facts by a process of literal repetition. With the discovery method, the number combinations were identified by the student through an active process of discovery or verification. The students were to discover their own errors and make the necessary corrections. McConnell perceived student discovery as apprehending and associating abstract symbols with number pictures which proved that certain addends produce certain sums. Teaching by the discovery method was to provide a meaningful environment for learning to take place. McConnell took much care to ascertain that the two groups did not have equal opportunities to learn the number facts. It was determined that students in the discovery groups worked on its learning materials 4 hours longer than the authoritative groups. The rote-meaningful factor seems to be more salient in this study than the discovery factor. McConnell (p. 26) states, "This experiment in the relative efficiency of two methods of learning the number facts is one of meaningful development versus sheer repetition."

Fourteen tests plus a pupil questionnaire were administered during the experiment. Seven interpolated and seven posttests were used to evaluate the instruction. The interpolated tests were given several times during instruction and consisted of achievement and transfer tests. The final tests included tests of transfer to untaught facts and tests of retention of addition and subtraction skills. All tests had high reliability coefficients; many of the tests were speed tests. The data were analyzed by using difference scores and standard errors of the differences.

McConnell concluded that the discovery group was superior on 10 of the 14 tests, which included tests of transfer to untaught number combinations and retention of addition and subtraction skills. The authoritative group was superior on tests requiring speed and accuracy and excelled on immediate and automatic responses to the number facts as measured by tests with limited administrative time and speed instructions, whereas the discovery group excelled on tests which put a premium on deliberate and thoughtful responses and those with generous administrative time.

This study was poorly controlled, and it is difficult to determine what caused the claimed superiority of the discovery method of teaching, the discovery techniques or the meaningful factor. At most, these findings suggest advantages for a meaningful-discovery method as opposed to an abstract-rote method for teaching addition and subtraction facts to second-grade students in terms of retention of computation skills and transfer to untaught facts.

Thiele (1938) conducted a 19-weeks study comparing the drill and generalization methods for teaching 100 addition facts to 512 second-grade students from nine schools. The addition facts involved only one-digit addends. The generalization method emphasized student discovery and use of relationships existing among addition combinations; students were encouraged to generalize whenever possible. Combinations were grouped according to unifying structure, and it was hoped that the student would attempt to generalize after working examples using visual aids and models. Exercises were used for review and practice, and reference was made to generalizations whenever possible.

For the drill method, the difficult combinations were presented in a random order to insure that patterns and relationships did not occur; this experimental bias favored the generalization method. Students were allowed to use concrete models, but only to verify the number combinations. Frequent use was made of a combination chart for checking answers. Memorization of the facts was sought by repetitive drill and practice. Little control was placed on the teacher variable. Each experimental group teacher was given a set of instructions and visited twice during the experiment.

Two tests were administered immediately after the instruction period, a 100-item addition-facts (retention) test and a 30-item transfer test. The addition-facts test was administered as both a pretest and a posttest to obtain a measure of growth in the learning of the 100 addition facts. The transfer test employed combinations with sums larger than 20. Concerning the transfer test, Thiele stated:

> Obviously a test which employs two-digit problems is not a transfer of training test for the pupils who perceived number generalizations in their study of the basic addition facts. It is rather a test which measures the ability of these pupils to extend generalizations to number situations involving larger numbers. For the pupils who did not perceive useful generalizations it does seem to test transfer of training. This shortcoming is recognized, but for the want of a better name the test is designated as, "Transfer of Training Test." (p. 52)

Both tests were administered using flash cards; four seconds were allowed for each card.

The pretest scores for both treatment groups were pooled for schools exposed to the same treatment and used to demonstrate that the two experimental groups were closely matched. The validity of Thiele's demonstration is questionable. He stated:

> The experimental groups in general were closely matched on the basis of the initial test. The difference of the means is 2.64, the reliability of which is 1.2 times the standard deviation of the difference, indicating a possible probability that a difference existed between the two groups. However, in all interpretations based upon the standard deviations of the distributions of the initial test scores, the fact that the distributions are somewhat skewed should be held in mind. Specifically, the distributions of the initial scores of both experimental groups are positively skewed to about the same degree. Therefore it would seem that tests of unreliability which are based upon the normal probability distribution should carry the same meaning for each of the skewed distributions. (p. 56)

This author does not understand why the reliability statistic should be invariant as to the distribution. The statistic 1.2 has no real interpretation since the parent population of its distribution is not known. Further, no justification was provided for pooling the data into two groups.

The posttest data were analyzed by visually comparing group means, by visually comparing standard deviations, and by computing a statistic found by dividing the difference of group means by the standard error of the difference of group means. This statistic was assumed to be normally distributed, a questionable assumption.

The results for the study as a whole, according to Thiele, indicated that the differences were decidedly in favor of the generalization method. Thiele stated:

> The superior results achieved by the generalization method pupils are in a very large measure due to the fact that they learned the so-called harder addition facts better than did the drill method pupils. (p. 76)

Due to a questionable data analysis and the confounding of the discovery and meaningful variables, no conclusive evidence supporting a superior teaching method for teaching certain number facts to second-grade students can be established. Thiele's use of nonrational drill for his non-discovery groups seriously handicapped their learning of the addition facts. His study is primarily one of contrasting rote and meaningful learning.

Fullerton (1955) compared the effectiveness of two methods of teaching multiplication facts to third-grade students. The two methods were the prescribed developmental method and the prescribed conventional method. The prescribed developmental method was described as

> . . . an inductive method characterized by pupil development of the multiplication facts from word problems by such means as counting, making marks, drawing pictures and diagrams, adding using the number line, and sooner or later multiplying. Each pupil is permitted and encouraged to work at the level of maturity appropriate for him, drawing upon his arithmetic background. Through this experience he sees the need for learning the multiplication facts and identifies multiplication as the most efficient process for solution of the problems. Throughout the period of instruction pupils were frequently called upon to show or prove that their solutions were correct. In

> addition to the number line, a counting chart and basic product cards were used in studying the number facts. Other aids employed in learning the facts were self-corrected tests and a definite study procedure, "How to Learn Facts." (pp. 168-69)

The prescribed conventional method

> . . . provided for virtually no active participation upon the part of the pupils in the development of the multiplication facts. Multiplication was immediately identified as the process to be learned and was introduced through addition examples and pictures illustrating the facts printed in the lesson material. . . . Less mature methods such as counting, additing making marks, and drawing pictures and diagrams were not permitted. No specific study suggestions were given. (pp. 169-70)

The instructional period consisted of eight forty-minute lessons. The multiplication facts taught were those with 2, 3, 4, and 5 as multiplicands and numbers 2 through 9 (inclusive) as multipliers.

Thirty third-grade classes from two school systems were involved in the study. The study was conducted in two parts. The first part contained 28 of the thirty classes, while the second part used the remaining two classes. Using class means from a pretest, the classes in Part 1 of the experiment were assigned to levels and then randomly assigned to methods within levels. For each treatment, four classes were assigned to the high level and five classes were assigned to each of the average and low levels. Each teacher taught his class the assigned method. The class mean was used as the experimental unit in Part I of the experiment.

In Part II, each class was assigned to three levels and then split in half and assigned to the two treatments. Each teacher taught both methods assigned to her class. The student score was the experimental unit for Part II of experiment. Immediately after instruction, a posttest consisting of two parts was given. Part 1 was a test of recall and Part 2 was a transfer test. A parallel form of the posttest was given three and one-half weeks later. Reliability coefficients ranged from .83 to .91 for the measures. For each study, a 2x3 ANOVA model was used to analyze the data. For Part II, by using four separate significance tests, Fullerton obtained significant differences (.05 level) on all the criterion tests favoring the prescribed developmental method. In Part II of the experiment, the only significant finding found occurred in the first class; the prescribed developmental method was significantly superior (.05 level) on the immediate-transfer test. For the other criterion measures, the prescribed developmental method was superior, but not significantly (.05 level) superior to the prescribed conventional method.

Very little control was placed on the teacher variable other than giving detailed instructions to the teachers. It was reported that a number of teachers using the prescribed developmental method had difficulty adequately covering the lessons in 40-minute class periods. Certain evidence of contamination of treatment was suggested. For, Fullerton, in discussing teachers using the prescribed conventional method, stated:

> In their written comments relative to their experience with the experiment several of these teachers either expressed disappointment at being assigned to the prescribed conventional method or made some comment which indicated that they did not believe it to be as effective an approach to multiplication as they would use in their regular teaching. (p. 178)

Concerning the analysis of the data, several questionable procedures were followed. The use of the student score in Part II of the experiment is questionable since the learning was not individualized. Further, the F-test for class B in the second study for the delayed-transfer test had a significantly low F-value, indicating a violation to the ANOVA model. The computed F was .08, and the critical value for the left-hand 5 per cent region was .21. Since .08 is less than .21, the value of .08 is significantly small, indicating a violation of the model. Since no power levels were reported for either study, it is difficult to determine the significance of the findings, especially those in Part II of the experiment.

These findings present some support for the superiority of the inductive-discovery for retention and transfer of learned multiplication facts. The use of intact classes, a questionable control of the teacher variable, and evidence of experimental bias concerning the meaningful factor prevent any conclusive inferences to be drawn.

Norman (1955) conducted a study comparing three methods (textbook, conventional, and developmental) of teaching basic division facts to 24 classes of third-grade children. The three methods varied the emphasis placed on making the lessons meaningful for the student. Teachers using the textbook approach were asked to teach following their usual teaching procedures; relatively little control was applied to this method other than having to cover certain textbook pages each day. Special materials, selected to be representative of those widely used for instruction in arithmetic, were supplied to the pupils in the conventional class. These material used story settings and problems to introduce the division facts. By liberal use of problems and examples, the pupils learned by repetition. Generalizations were always stated for the students. No instructional aids other than the conventional material were used to make learning meaningful.

The developmental method involved the use of instructional aids, such as the number line, drawings, and counters. Students were encouraged to make generalizations. After students had some experience using several different methods to obtain answers to verbal problems, the methods were compared in order to determine the best method to use. Students were then introduced to the conventional notation for division problems. The experimental instruction involved eight forty-minute lessons. One test, divided into two subtests, was given as a pretest, a posttest, and as a two-week retention test. Part I measured simple recall and Part II was a transfer test to untaught division facts. On both parts, the majority of the items were given orally by the teacher; 15 seconds were allowed for each question. Split-half reliability coefficients were reported and ranged from .88 to .94, with a coefficient of .90 for the total test.

Of the 24 classes used in the experiment, six classes (N = 164) were assigned the conventional treatment, six (N = 155) to the developmental treatment, and 12 classes (N = 323) to the textbook treatment.

The experimental unit for the study was the class median. The data were analyzed by using an ANCOVA model. It is not clear just what the covariate was, presumably the pretest score. Six F-tests were employed, and a significant F-value was followed by pairwise t-tests to detect pairwise differences. No power levels were reported for the significance tests.

Norman reported that (1) there were no significant differences among the three groups in initial achievement; (2) the developmental treatment was superior to the textbook treatment on Part I, Part II, and total delayed-recall tests; (3) the developmental method was superior to the conventional method on both parts of the delayed recall test, but not on the total test; and (4) the conventional method was superior to the textbook method on Part I and the total test of delayed recall.

The use of the group median as the experimental unit is questionable. Dixon and Massey (1969, p. 130) point out that for normal populations, the distribution of class means has a smaller variance than the variance of the distribution of class medians. This implies that a single observed mean has a greater chance of being close to the true (population) mean than does a single observed median.

The experimental design contains several aspects of questionable validity. Using only one form of a test as both a pretest and a posttest over a period of a month, may not have controlled for the carry-over effects of learning. Further, the pretest may have sensitized the learners to the treatment, thus confounding these two variables. The meaningful factor possibly introduced experimental bias in favor of the developmental method. The extent of control of the teacher variable is questionable; no rigid procedures were followed to insure that the teacher was faithful to his assigned method. Oral examinations in group format may also have contaminated the experiment.

The use of pairwise t-tests is an inappropriate probability model to follow for post hoc comparisons and leads to inflated and uncertain alpha levels. Furthermore, the experimentwise error rate, based on the six F-tests, cannot be determined since the criterion variables are undoubtedly correlated.

Due to questionable methodology, Norman's findings should only be considered as tentative. The developmental method suggests superiority in terms of delayed retention and delayed transfer.

In a well controlled study, Worthen (1968) compared discovery and expository task presentations which differed primarily in terms of sequence characteristics. The study involved 538 fifth and sixth-grade students from eight different elementary schools; a total of 19 different teachers and classes were involved. In each of the eight schools, two classes were taught arithmetic by the same teacher, one class by the discovery treatment and one class by the expository treatment. The students were taught number concepts and principles for a period of six weeks by teachers using prepared, textlike materials. The subject matter consisted of (1) notation, addition and multiplication of integers, (2) the distributive principle, and (3) multiplication and division of numbers in exponential notation.

In the discovery method, the student was presented with an ordered, structured sequence of examples of a generalization. No explanations of the examples were given. Verbalization of discoveries by the teacher was delayed until the end of the instructional sequence.

In the expository method, the verbalization of the required concept or generalization was the initial step followed by examples only after the initial verbal and symbolic presentation. In addition to the sequencing of subject matter, five other aspects of teaching behavior were controlled. These include the following: (1) interjection of teacher knowledge, (2) introduction of the generalization, (3) method of answering questions, and (4) control of pupil interaction, and (5) method of eliminating false concepts. With regard to (4), student interaction was encouraged in the expository treatment, but discouraged in the discovery treatment. This procedure could have resulted in experimental bias favoring the expository method.

Two instruments were used to assess the degree to which teachers adhered to the prescribed teaching model assigned to them. These two methods were pupil perception of teaching behavior and observer ratings of teaching behavior. Analyses of both instruments indicated teacher fidelity to methods.

Nine measures were developed to compare the instructional strategies: a concept knowledge test, two concept-retention tests (5 and 11 weeks after instruction), a concept-transfer test, a negative-concept-transfer test, two heuristic-transfer tests (oral and written), and two attitude scales. The negative-transfer test was included in order to assess the student's tendency to overgeneralize the principles to inappropriate situations. Test-retest and parallel-forms reliability coefficients ranged from .69 to .87, with a mean of .72.

Ten two-way, teacher-by-treatment, ANCOVA designs were employed to analyze the data. I.Q., arithmetic computation, and arithmetic problem-solving were used as constant covariates in the analysis of each dependent variable. In addition, whenever appropriate, pretest scores were also used as a covariate. Posttest scores on the concept knowledge test were used as an additional covariate in the analysis of the concept retention and transfer tests. The student score was used as the experimental unit in all analyses.

Worthen concluded that the expository group was superior on initial learning, while the discovery groups were superior on the retention and transfer of heuristics tests. The discovery group was slightly superior (.08 level) to the expository group on a transfer of concepts test. Significant F-ratios for between-teacher effects and teacher-by-treatment interaction were yielded by the analysis of each criterion measure. This could mean that teacher personality is an important variable in methodology studies.

Worthen's analysis of his data is suspect. Since student interaction was present in the study, the use of the student score, as opposed to the class mean, is suspect. Further, the posttest scores on the concept knowlege test were influenced by the treatments and should not have been used as a covariate.

Reanalyses of Worthen's data are provided by Worthen and Collins (1971).

> . . . test scores from both transfer tests and both retention were reanalyzed using the design of the original study but deleting the Concept Knowledge posttest as a covariate. This reanalysis yielded no significant difference between treatments on any transfer or retention test. (p. 15)

Furthermore, when the data were reanalyzed using the more conservative experimental unit, the class mean, the significant differences on the transfer and retention tests again vanished. Viewed collectively, these reanalyses suggest that the results of Worthen's study should be considered with caution.

Keurst and Martin (1968) conducted a study comparing expository and discovery methods of instruction. In this study, 26 fourth graders were to find the sum of an arithmetic progression with an odd number of terms by multiplying the middle term by the number of terms in the sequence. The two experimental groups, consisting of 13 students each, spent 10 minutes a day for five days solving problems. The progressions were all arranged in ascending order. The rote learning group was given the rule and the discovery group was taught by an improvised "see-saw" technique that the average identified the balance point in hopes that they would discover the rule.

After the training period, both groups were given an examination that consisted of 10 new problems, three of which were not arithmetic progressions and two of which were arranged in a descending order of magnitude. The experimenters make no mention of content validity or reliability coefficients for the test. A t-test analysis indicated that the rote learning group was significantly (.01 level) superior on a test of immediate learning and on a similar test measuring retention given three weeks later. The experimenters give no standard deviations for either group, nor do they mention how the teacher variable was controlled. It may have been the case that the rote learning group discovered the rule during the training period, thus comparing two discovery groups on the criterion measures. The study appears to have been poorly conducted, and is too poorly reported to draw any valid inferences from it.

Werdelin (1968) compared three programed methods (rule-example, example-rule-example, and example) for teaching the left distributive principle of multiplication over addition to seven classes containing 178 sixth-grade Swedish students. Students were randomly assigned to the three treatments.

Students in the rule group were given the rule accompanied by twelve solved examples. They were then allowed to practice the rule on another 78 examples. Students in the rule-example-rule group were first exposed to 74 examples which they were told to solve. They were then given the rule along with three solved examples which were followed by 13 more practice examples. Students in the example group were given 90 examples to solve with no explanations given. The examples ranged in difficulty, e.g., 20x2 + 20x8 =, 32x2 + 32x48 =, and 94x36.95 + 94x23.05=.

To evaluate the effectiveness of the three programs, four posttests of eight items each were given immediately after the learning session. Werdelin does not indicate the length of the instructional period. Each test took six minutes. It was not reported whether all the tests were given in one class period or not. No validity or reliability information concerning the tests were reported. Test 1 contained items of the same type as those appearing in the lesson. Test 2 contained items dealing with the right distributive principle of multiplication over addition. Items on Test 3 dealt with the left distributive principle of multiplication over addition applied to three terms, while Test 4 contained items illustrating the distributive principle of multiplication over subtraction. To measure retention, four parallel tests were administered two weeks later.

Werdelin reports that the data were analyzed using the common parametric test for differences between proportions, the Kolmogorov-Smirnov two-sample test, and the Sign test. No standard deviations were reported; hence, it is impossible to determine whether his findings are in keeping with the data. He concluded that the rule-example group was superior to the other two groups on tests of immediate learning, while the example (discovery) group was superior to the other two groups on the delayed retention and transfer tests.

It is possible that the students in the example-rule-example group may have suffered from the effects of retroactive inhibition. Werdelin felt that students who were given the rule first may have concentrated on the syntactic dimension of the rule whereas students who discovered the rule from examples may have understood the semantics of the rule, thus leading to superior delayed retention and transfer (p. 247).

A study by Barrish (1970) was undertaken to determine the relationship between two levels of divergent thinking and the effectiveness of inductive-discovery and deductive-expository teaching strategies. Students in grades four, five, and six from an open-classroom school were stratified by grade level, I.Q. score, and two levels of divergent thinking and were randomly assigned to four treatment groups, two expository and two discovery. Each group had 32 students, 16 high-divergent and 16 low-divergent thinkers as judged by the Torrance Tests of Creative Thinking.

Four female teachers taught both treatment groups. After ten days of a 20-day instructional period, the teachers changed groups and instructional strategies. According to Barrish, the students were accustomed to team-teaching procedures. Each teacher had less than three years teaching experience at this level, with two teachers having less than two years experience. Only one teacher had experience in developing inductive teaching strategies.

Provisions were made for controlling the teacher variable. Teachers were trained in a five-meeting training session concerning the instructional programs. Each teacher was observed frequently using protocol consisting of both objective and subjective evaluative criteria by personnel who were not involved in the research. Teacher effects within treatments were determined by a sign test based on binomial probabilities.

The content for the experimental lessons consisted of graph; areas of squares, triangles, and rectangles; volumes; angles; and metric system topics.

The expository strategy first presented a generalization of a principle followed by showing examples, answering questions, and clarifying and reiterating the principle. Drill and concrete manipulation of models were used where feasible. Student feedback was encouraged.

The inductive-discovery strategy elicited discovery through discussion by giving examples and answering student questions. Leading questions were sometimes asked by the teacher. The discovered generalization was always restated. Drill and practice with concrete manipulation of models were used wherever feasible.

A test of acquisition was given after the instructional period followed by an equivalent test of retention given 20 school-days later. The 35-item acquisition test was divided into two subtests: a 25-item test of recall and a 10-item transfer test involving independent thinking and applications to novel situations. No validity or reliability data were reported for either measure.

Barrish reported that some minor problems were encountered during the course of the experiment. Several teachers were absent during the course of the experiment and were replaced by trained substitutes. There was some difficulty encountered by several teachers in teaching the operations in modular arithmetic. Further, two teachers distributed the wrong work-sheets by accident one day, and a discovery teacher taught one of the examples deductively. Two teachers used questioning techniques extensively in their deductive teaching. The four teachers were observed a total of 56 times during the experiment, no teacher being observed less than 12 times. Of this, Barrish (p. 61) stated, " . . . all but five clearly indicated that the teachers were closely following the strategy of teaching to which they were assigned."

Analysis of covariance and regression analysis were used to analyze the data. Several questionable procedures were followed. First, the student score was used as the experimental unit for the ANCOVA rather than the class mean. Since the instruction was not individualized and student feedback was encouraged, the class mean appears to be the appropriate experimental unit for this analysis. Secondly, the data from the four classes were pooled into two groups (discovery and expository) for analysis without some justification for this procedure.

The results, according to Barrish, indicated that the expository group was significantly (.01 level) superior to the discovery group on the acquisition test and on the recall subtest. For the retention test, the results favored the expository group for the recall subtest at the .05 level of significance. No other significant results were reported.

The sign tests for teacher effects showed that in both the inductive-discovery and deductive-expository groups, differences in teacher efficacy were far from significant at the .05 level of significance. For these tests, student scores on the Torrance Test were paired by teacher team within each treatment. Their scores on the test of acquisition were then compared.

Barrish's failure to report reliability data for his achievement measures and his questionable use of the student score, rather than the class mean, as the experimental unit make it risky to draw any conclusive evidence from this study. By having two female teachers instruct each treatment group, the treatment variable was confounded. This author does not understand how the results of the sign tests indicated that the teacher effect was controlled for. Furthermore, the sex of the teacher may be an important variable in studies of this nature. The novelty effect of having different teachers teach each group may have differentially effected the treatments even though students were accustomed to team teaching.

Scott (1970) conducted two experiments with sixth-grade students comparing the effects of discovery and expository methods of programing on the immediate acquisition, retention, and transfer. The first experiment was concerned with short- and long-term retention, and the second experiment was concerned with immediate acquisition and transfer. The expository method was a rule-example method, and the discovery method was inductive.

The subject matter consisted of two units, one on the triangle and one on the quadrilateral. The triangle unit dealt with such concepts as equilateral triangle, isosceles triangle, and right triangle. The quadrilateral unit contained the concepts of quadrilateral, rhombus, parallelogram, and trapezoid. These particular concepts were chosen because of their unfamiliarity to the learners.

Under both methods, students were shown a series of six examples, four positive and two negative (+, -, +, -, +, +). Discovery students were asked how these examples were alike and how they were different and then given the name of the concept. Students using the expository method were first given the name of the concept followed by the examples with the relevant attributes pointed out. The students studied the prepared lessons for four days in the first experiment and five days in the second experiment.

Two-hundred and fifty-six six-grade students from six schools were involved in the study. Students were randomly assigned to 15 different treatment groups within each school.

Nine groups were involved in the first experiment and six in the second experiment. In the first experiment, three groups were assigned to each of the treatments, expository, discovery, and control. The unit on quadrilaterals was used as the subject matter in the first experiment. One group took the posttest one day after the treatment, one group took the posttest 11 days after the treatment, and one group took the posttest 21 days after the treatment. Each group took only one posttest.

Two tests, Test E and Test Q, were used in both experiments. Test E was a test embedded in the lessons and was a production test which required the student to produce a word, definition, or to complete a figure. Test Q, a multiple choice test containing 28 items, was a parallel form of Test E and required recognition. A reliability coefficient of .83 was reported for test Q.

In the second experiment, both the triangle and quadrilateral units were used. Six treatments were formed: triangle lesson-discovery mode, quadrilateral lesson-discovery mode; triangle lesson-discovery mode, quadrilateral lesson-expository mode; triangle lesson-expository mode, quadrilateral lesson-discovery mode; triangle lesson-expository mode, quadrilateral lesson-expository mode; quadrilateral lesson-discovery mode; and quadrilateral lesson-expository mode. After finishing the quadrilateral lesson, each group took Test Q.

It is not clear how transfer was measured. Presumably, it was a gain score computed with reference to the groups experiencing the quadrilateral lesson only. Scott did not define transfer or how it was to be measured.

For both experiments, the data were analyzed by using ANOVA and ANCOVA models. It is not clear how the treatments were compared with the control groups; in the factorial design, the treatment factor had only two levels, discovery and expository. When using the ANCOVA model, Test E was used as a covariate, a questionable move since it was influenced by the treatment. Separate ANOVA analyses were conducted and verified the results of the ANCOVA tests. Replicating analyses in this fashion is not desirable since it inflates the experimentwise error rate as well as leading to confusing results. No power levels were reported for either experiment.

Scott concluded that the method of presentation did not differentially affect immediate acquisition or transfer, but did differentially affect retention of the material in favor of the discovery method. For overall retention intervals, the expository method was superior to the discovery method, but interaction of treatment and retention was significant and indicated that the expository method effected a net loss over time and the discovery method resulted in a net gain over time. Since Scott does not define *retention*, *transfer*, or how they are to be measured, it is difficult to interpret his results, especially for transfer.

Scott's school-by-treatment-by-retention-interval factorial design for the first experiment controlled for any carry-over effects of being exposed to repeated measures. It is conceivable that repeated measures could provide an opportunity for the student to acquire a concept because of the wording or frequency of seeing the test questions.

Scott's experiments were well planned and provided for adequate control of many variables that could have contaminated the study. It was not reported if any effort was made to control for the novelty effects of the treatments, including the effects of the programed mode of teaching.

This experiment provides some strong evidence for the superiority of an inductive-discovery method, compared with an expository strategy, for teaching certain quadrilateral concepts to sixth-grade students.

Cooke (1971) conducted a study to determine the effects of three strategies on the conceptualization of linear block design sequences in first-grade children. Forty-eight students from five schools in a public school system were involved in the main study. All students had an I.Q. greater than or equal to 95. Students were matched according to sex, I.Q., and age and were randomly assigned to the three treatment groups.

The stimulus materials consisted of 75 attribute blocks, each invested with four attributes: color (red, yellow, and blue), size (large and small), thickness (thick and thin), and shape (circle, square, and triangle). Individually, the students were shown five pairs of block designs, one pair at a time. Each pair of designs contained two horizontal sequences of nine blocks each and equally spaced. The first two pairs of designs were arranged according to shape, the third and fourth pairs according to color, and the fifth according to shape and color. Each design had an organizing principle associated with it. Concerning the learning tasks, Cooke stated:

> . . . The learning task consisted of each subject's reproducing from memory, five linear series of 9 blocks each arranged according to the embedded principle. The subjects were presented with five pairs of designs. The subject modeled the first design of the pair in accordance with the sequence of interaction assigned to his treatment group. This manipulative process was designated to allow the child to build basic information about the design and was called the association task. The subject was then shown the second design of the pair for twelve seconds. The second design had the same embedded principle and varied the same attribute, however, it had a different arrangement and was called the conceptual task. After the subject studied the conceptual design for twelve seconds, it was removed and he attempted to select the correct blocks from his pile of seventy-five and replicate the conceptual design on his tray. After completing the first pair of designs the subject was taken through the remaining four pairs of designs in a similar manner. (p. 35)

The three strategies were identified as rote, principle, and discovery methods. Students in the principle-learning group were told the organizing principle which related the individual elements to the sequence pattern. In the rote-learning group, the student's attention was focused on singular elements and pieces of information; they were given attribute cues, but not the organizing principle. The guided-discovery strategy employed questioning in such a way that the student was led to formulate his own conceptualization concerning the relatedness of singular elements. Questions were directed towards the attributes and ordering of the design.

Cooke does not indicate whether the treatments were randomized according to time of day they were conducted.

In order to control the teacher variable as much as possible, the examiner had the instructional sequence in front of him on typed cards.

A maximum score of 45 could be scored for each conceptual sequence, one point for each of the four correct attributes and one point for the correct position in the sequence. The total score for the five conceptual sequences became the basis for the analysis. After one and six weeks and without further instruction, the students were retested using the same five pairs of block designs.

The data were analyzed by employing a 2x3x3, sex-by-treatment-by-measure, ANOVA design with repeated measures on the third factor. The main effects for treatment and measures were found to be significant at the .01 and .05 levels, respectively. Using Scheffe's test, it was found that there were significant mean differences in performance for all groups from the initial test to the first retention test and from the initial test to the six-week retention test. For each measure, the group means were ranked with the rote-group mean less than the principle-group mean, and the principle-group mean less than the discovery-group mean. The mean scores of both principle- and rote-groups diminished from the end of the first week to the end of the sixth week; however, the mean scores of the discovery group continued to increase through the final test. Using Scheffe's test to test for significant differences in the treatment effect, it was found that the discovery method was significantly (.01 level) superior to the rote method and the discovery group was significantly (.05 level) superior to the principle-learning method. Cooke concluded:

> In tasks where a specific principle is to be learned, a guided discovery technique produces better initial results, which are sustained and improved over long periods of time, then are principle or rote strategies. . . . Guided discovery strategies provide some of the efficiency of principle learning (reducing the time it takes the learner to find the answer) while at the same time allowing the learner to develop some of his own initiative for learning. (p. 72)

This experiment has several weaknesses. No tests for homogeneity and compound symmetry of the covariance matrices were reported; both are required as preliminary assumptions for a repeated-measures design. No reliability information is given for the criterion measure. A stability measure could have been obtained by correlating the initial achievement and retention measures. An alpha level of .05 was set as the experimental standard, but no power levels were reported for the significance tests.

This study suggests that a discovery method is more effective than an expository method for teaching first-grade children certain concepts and having them retained over a six-weeks period. Without reliability estimates, no strong inferences can be drawn from the data. As a result, his findings can only be considered as tentative.

Murdoch (1971) studied the relative effectiveness of inductive-discovery and deductive-expository methods of teaching certain concepts to fourth-grade students. With each presentation method, two methods of stimuli presentation were compared; the stimuli were presented all at once or one at a time and then removed.

Three single-attribute concepts were taught, each in a separate lesson. The concepts were number series, topsy words, and alpha designs. The number series task required the student to identify arithmetic progressions of five terms. The topsy-words task required the students to identify two-vowel non-words which would become words by transposing both vowels. The alpha-designs task required the students to identify geometric designs which contained at least one straight line.

In all treatments, 20 stimuli were shown, 10 positive and 10 negative examples, by the classroom teacher. Murdoch does not indicate whether the positive and negative examples were sequenced in any particular order. After a stimulus had been presented and a response (presumably nonverbal) was made by the class, the word yes or no was made visible to the students. Murdoch does not indicate the length of the instructional periods.

Immediately following each concept presentation period, the students were presented with a yes-no, 15-item, concept learning test. The tests were scored by subtracting the total number incorrect responses from the total number of correct responses. This was done to correct for guessing. The same three tests were given four weeks later as a retention measure. No reliability coefficients were reported.

The accessible population consisted of 287 fourth-grade students from 12 classes in four elementary schools. The 12 classrooms were randomly divided into four groups of three. Within each classroom, students were ranked by I.Q. score from highest to lowest; three ability level groups were then formed. Four students were then randomly selected from each I.Q. level in each class to obtain the random sample, 12 students from each class resulting in a total of 144 students.

The treatments were administered by each regular teacher to all students (22-28) within the class; only the scores of the students involved in the sample were used in the analysis. Three classrooms were randomly assigned to each of the four treatment combinations obtained by crossing the two methods of instruction with the two methods of presenting the stimuli.

The data were analyzed by using a four-factor ANOVA design with repeated measures (concept tests) on the last factor. The student score was used as the experimental unit. Murdoch does not indicate in his report whether or not there was any student interaction during the instructional periods; so one can question the applicability of the student score as the experimental unit.

It was concluded that all main effects of methods, stimuli presentation, ability level, and concept tasks were significant for learning and retention measures. The deductive-expository method was significantly (.01 level) superior to the inductive-discovery method on all learning and retention measures. No difference was found regarding the presentation format of the stimuli upon the amount of learning or retention. The high-ability level had higher learning and retention scores than the low level with the middle level not different from either.

This study has several weaknesses. An inadequate report of the treatment procedures makes it difficult to determine if any intervening variables contaminated the treatments. The 15-item tests, corrected for guessing, appear to be too short to reflect true achievement or retention. The correction factor for guessing penalizes the student who does not guess. Reliability coefficients should have been reported. The topsy-words task assumed that the students were competent spellers, a questionable assumption. Otherwise, the concept was not appropriate for the fourth-grade level.

This study offers some support for the superiority of an expository method of teaching certain one-attribute concepts to fourth-grade students as judged by achievement and retention measures. But, without reliability estimates for the criterion measures, no conclusive evidence concerning the dependent variables can be drawn.

Nichols (1971) compared two methods of teaching multiplication and division facts to 267 third-grade students from ten classes in six different elementary schools. Treatment A employed manipulative materials and guided-discovery techniques. The manipulative materials consisted of 5/16-inch metal nuts and felt mats. Children worked in pairs to discover the number facts. Multiplication was taught by identifying X with sets of or groups of. That is, 2 x 3 = 6 was read as 2 sets of 3 equals 6. Also, students were taught that ÷ means how many groups of. That is, 6 ÷ 2 was read as six has how many sets (groups) of two?

Treatment B employed abstract and semi-concrete materials combined with teacher explanations and exposition. Children were not permitted to use concrete objects; they could look at pictures, draw diagrams, or observe the teacher manipulating concrete aids. Fifteen 45-minute instructional periods were used in the study. Instruction was in the regular classrooms and at the regular time used for studying arithmetic.

Teachers were randomly assigned to 10 intact classrooms, five classes of treatment A and five classes of treatment B. The students in the two treatments were not matched, according to pretests results. Concerning this, Nichols states:

> In summary, pretest scores for the subjects assigned to treatment B showed them to have a slightly lower chronological age, significantly higher mean I.Q. scores on eight subtests, and significantly higher mean subtest scores on five subtests of the test of General Arithmetic Knowledge. Treatment group A showed significantly higher ability on one subtest score, computation in multiplication. Treatment group B also showed significantly higher scores in tests of understanding arithmetic and in tests of attitudes toward arithmetic. (p. 41)

Four tests were administered as posttests and retention tests after the learning period to measure general arithmetic knowledge, attitude toward arithmetic, and understanding of arithmetic. No validity or reliability data were reported for any of the four tests.

ANOVA and ANCOVA models were used to analyze the data. For each measure, hypotheses were concerned with posttest mean gains, difference in durability of mean gains between groups, differences in posttest mean gains for students having an I.Q. of 105 or lower, and differences in durability of mean gains for students having an I.Q. of 105 or lower. Sixteen hypotheses were tested using 28 F-tests.

Nichols' analyses of the data are questionable. Gain scores generally are less reliable than nongain scores. Not knowing the reliabilities of the measures involved makes it risky to put much faith in gain scores. Twenty-eight F-tests admit "probability pyramiding"; one can not be sure of just how much greater than .05 the experimentwise alpha level is. An insufficient and inadequate report of the statistical analyses was made; ANOVA and ANCOVA summary tables should have been provided so one could determine if the reported findings are in keeping with the data. Without justification, the five groups of data within each treatment were pooled to form a single group; this is a questionable procedure. The student score was used as the experimental unit rather than the class mean. Further, by using gain scores for students with I.Q. lower than or equal to 105, regression towards the mean becomes a factor in the analysis, possibly confounding the results.

The experimental design fails to adequately control the experiment. There was questionable control placed on the teacher variable. Teachers in both treatments were supplied with plans for instruction, but no checks were reported to ascertain whether the teachers were faithful to their assigned methods. Using intact groups makes it impossible to control the experimental errors by using covariance analysis. The salient factor in this study appears to be the type of learning aids used rather than the discovery variable. Failure to randomly assign the treatments to room and time period may have introduced further confounding.

Nichols concluded, in every instance, that the guided-discovery treatment was significantly superior to the expository method. This experiment is poorly controlled, and a questionable analysis of the data make it impossible to determine if Nichols' findings are in accord with the data. The raw data were not included in the report.

In a 31-weeks study, Olander and Robertson (1973) compared discovery and expository methods of teaching fourth-grade students mathematics with emphasis upon principles and relationships. Three hundred seventy-four fourth-grade students and 13 teachers from a certain school district were involved in the study. Seven classes with 190 pupils were involved in the discovery treatment, and nine classes with 184 pupils were involved in the expository treatment. No school building had classes assigned to both treatments; this was done to minimize conferences among teachers that could result in contamination of the data. Buildings were assigned at random to treatments. It was not reported whether students were randomly assigned to classes or whether intact classes were used.

In the discovery treatment, teachers promoted search behavior of students generally through problem situations. A Socratic form of questioning was used by the teachers, and concrete models and representative materials were used to facilitate discovery. Pupils were discouraged from helping others or "giving away" the right answers.

In the expository treatment, students learned through explanations by the teacher or through those in the textbook. The students were encouraged to seek help and to share ideas about a problem with their peers.

Measures were taken to adequately control the teacher variable. The teachers were all rated on a scale measuring the degree to which each succeeded in using the method assigned to them. Mean rating scores were computed for each teacher on the basis of ten classroom observations. The analysis of these scores indicated that the two treatment groups differed significantly in regard to teaching practices.

The relative effectiveness of the instructional programs was evaluated by employing the following criterion measures: The Stanford Achievement Test, a 150-item test designed to measure understandings of mathematical principles and relationships, and an attitude-toward-mathematics scale. The Stanford Achievement Test contained subtests measuring pupil performance of computation, concepts, and applications. The principle and relationships test had a reliability coefficient of .94. No reliability information was reported for The Stanford Achievement Test. Both the Stanford and principles and relationships tests were given as pretests, posttests, and five-week retention tests.

The pupil-outcome (cognitive) data were analyzed by using eight one-way ANCOVA designs. Pretest scores were used as covariates for posttest and retention scores. The student score was used as the experimental unit in the analyses of the principles and relationships test scores. The experimental unit for the analyses of the Stanford Achievement Test scores could not be determined from the experimenters' report.

Olander and Robertson reported several significant findings. On both the posttest and retention test, students taught by the expository method did significantly (.05 level) better than students taught by the expository method on the computation tests. There were no significant differences on posttests measuring concepts, applications, or principles and relationships; however, students taught by the discovery method were significantly (.05 level) superior on the retention test of applications. Measures of student attitudes revealed that pupils taught by the discovery approach improved their attitudes towards mathematics significantly (.05 level) more than those students taught by the expository method.

This study contains several design problems. No report was made of controlling the effects of pretest sensitization of students to treatments. Experimental bias favoring the expository treatment may have resulted from the experimenters' control of pupil interaction; student interaction was encouraged for the expository treatment but not for the discovery treatment. The experimenters report (p. 35) that greater use was made of learning aids by the discovery students since the expository students depended largely on the textbook. This treatment bias may have also confounded the study.

Several questionable procedures were employed in analyzing the data. Since the treatments were not individualized, the use of the student score as an experimental unit of analysis is suspect. There was, clearly, interaction among the students in the expository classes. Further, the pooling of data into two treatment groups for analyzing the mathematical principles and relationships test scores was not justified and is open to criticism. The experimenters report no preliminary checks to ascertain that the assumptions underlying the ANCOVA model had been met.

From Olander and Robertson's report, it is difficult, if not impossible, to determine whether their findings are in keeping with the data. Neither degrees of freedom nor error terms were reported for the analyses of the pupil outcome scores. Furthermore, no mention was made of checking for homogeneity of regression slopes, a prerequisite assumption for the ANCOVA.

Due to confounding of variables in this study, any reported differences cannot solely be attributed to the discovery variable. This study suggests that both treatments are effective for teaching certain mathematical content to fourth-grade students.

This experimental design appears to have been well planned and controls for many independent variables, including the teacher variable and the effects of novelty.

Olander and Robertson's findings suggest that expository students are superior for retaining computational skills, whereas the discovery students are superior for retaining the ability to apply certain mathematical content.

In summary, of the 14 studies investigating retention of mathematics at the elementary level and reporting significant findings, only the studies by Scott (1970), Cooke (1971), Murdoch (1971), and Olander and Robertson (1973) offer supportive evidence for the superiority of certain teaching strategies. All four of these studies appeared, from the written reports, to have been well controlled. The studies by Scott (1970) and Cooke (1971) favor the discovery method for teaching certain geometrical or classificational concepts as measured by retention tests. The retention interval for Scott's experiment was three weeks, while Cooke's retention interval was six weeks. The study by Murdoch (1971), on the other hand, suggests that the expository method is superior for teaching certain single attribute concepts as measured by a four-week retention test. These three studies, taken as a whole, suggest that simple concepts are more easily retained when taught by expository methods, while complex concepts (judged according to the number of defining attributes) are more easily retained when taught by discovery methods. The study by Olander and Robertson (1973) seems to imply that students exposed to an expository technique are better able to retain certain computational skills as measured by a five-week retention test; whereas, students experiencing a discovery treatment appear to better retain their ability to apply mathematical knowledge as measured by the same retention test. Olander and Robertson did not indicate what kinds of knowledge were used in the applications. This author finds it difficult to distinguish, in this case, between retaining the ability to apply knowledge (retention) and the transfer of learning.

Junior high level (grades 7 - 9). Table 2 summarizes the findings concerning retention of mathematics at the junior high level as reported by the experimenters. Only two from a total of 12 studies yielded statistically significant findings concerning retention; both of these studies

## Table 2

## Method Favored as Determined by the Significant Findings Reported by Experimenters for Retention of Mathematics

Junior High Level

(Grades 7 - 9)

| Discovery | Expository | Neither |
|---|---|---|
| Neuhouser (1964 | | Bassler *et al.* (1971) |
| Sobel (1954) | | Brown (1969) |
| | | Eldredge (1965) |
| | | Hanson (1967) |
| | | Kuhfittig (1972) |
| | | May (1965) |
| | | Maynard (1969) |
| | | Meconi (1967) |
| | | Nelson and Frayer (1972) |
| | | Strickland (1968) |
| Total 2 | | 10 |
| Per cent* 17 | | 83 |

*Based on a total of 12 studies.

favored the discovery method. These studies were conducted by Sobel (1954) and Neuhouser (1964).

Sobel (1954) compared a deductive method and a non-verbalized, inductive method for teaching concepts and skills to ninth-grade algebra students. The deductive method was characterized as being abstract and verbalized with teacher exposition followed by practice, whereas the inductive method, through experiences involving applications, guided the students to discover concepts.

Fourteen intact classes (seven inductive and seven deductive) from seven high schools were used as the experimental population. Teachers were not randomly assigned to methods; instead, teachers were assigned to methods based on their style of teaching as judged by their department chairmen. Sobel states:

> No random assignment of teachers to methods was deemed feasible since the experimental method required considerably more effort on the part of the teacher than the control method and not every teacher was willing to participate to this extent. It was also felt that a teacher would be able to produce more valid results if the method of approach used was one in which there was interest and enthusiasm, and which corresponded more nearly with the method which the teacher normally used. (p. 20)

Each teacher using the inductive method was given a manual of instructions which outlined the specific material to be covered during the four-week study, but the teacher was free to develop each day's procedure in his own way. Three conferences with each experimental teacher were held prior and during the experiment to discuss problems and progress. No other procedures were taken to guarantee teacher fidelity to the inductive method.

The teachers using the deductive method were given instructions to have their students master a set of specific skills using the method found in the textbooks. Sobel states:

> . . . In each case, assurance was given by the department chairmen that the group of teachers was not actually teaching by the experimental method, although they naturally differed in the degree to which they followed the approach designated as the control method. (p. 21)

This lack of control of the treatment variable confounds the study, and the non-randomization of teachers and subjects limits the external validity of the study.

Prior to the experimental treatment, an inspection of the mean I.Q. for each class indicated that each class could be identified as being average (100) or high (110-115) in intelligence. Sobel does not indicate whether these levels are based on a single I.Q. measure or not. Nine average ability classes were identified along with five high ability classes.

A posttest, consisting of two parts, to measure concepts and skills was given to both groups at the conclusion of the experiment. A parallel form was given three months later as a retention test.

The data were analyzed with respect to the following classifications:

1. deductive and average I.Q. (4 classes)
2. deductive and high I.Q. (3 classes)
3. inductive and average I.Q. (5 classes)
4. inductive and high I.Q. (2 classes)

Sobel pooled the data within each of the four major subgroups after verifying that homogeneity of means and variances were satisfied at the .05 level of significance. Hartley's $F_{max}$ test was employed 20 times, four for the I.Q. measure, eight for the posttest, and eight for the retention test. Twenty one-way ANOVA tests were given to determine homogeneity of means, four for the I.Q. measure, eight for the posttest, and eight for the retention test. All but one of the tests yielded non-significant results. In testing for homogeneity of variance, it was found that the $F_{max}$ statistic for the deductive, average I.Q. subgroup was significant. Sobel attributes this to the low variance in I.Q. scores where an I.Q. measure for the students in one school was not available and the Otis Quick Scoring Test was administered at the start of the experiment. Sobel states:

> . . . It can be expected that a lower variance would result when all the members of a group take the same test, at the same time, and under the same conditions. (p. 30)

Of the 40 hypotheses, if independent, two would be expected to be rejected by chance alone. The .05 level seems to be too conservative a level since inequality of classes could seriously affect the validity of the analysis used. A more liberal level, in the order of .15-.25, would make discovery of inequality of classes more likely.

Three of the one-way ANOVA tests yielded F-ratios significantly less than one under the criteria established by Myers (1966, pp. 66-67).

> . . . The occurrence of F's so small that their reciprocals are significant or the occurrence of many F's less than one in a single analysis of variance merits further consideration. Such findings suggest that the model underlying the analysis of variance has in some way been violated.

This violation of the model makes any interpretation of the data suspect.

After pooling the data, the final analysis employed eight one-way ANOVA tests, four for the posttest and four for the retention test, at the .05 level of significance. For each criterion test, assuming independence, the probability of committing a type I error becomes .19 instead of the experimental standard of .05. ($.19 = 1-.95^4$) Two comparisons were made for each subtest (concepts and skills) for both the posttest and retention test. Four of these eight comparisons yielded significant differences favoring the brighter students who learned concepts and skills under the inductive method. Sobel concluded that brighter students profit in the learning of certain concepts and skills from an inductive approach as opposed to a deductive approach to learning.

The experimental design and statistical analyses used in the study make it extremely difficult, if not impossible, to draw any meaningful conclusions.

Neuhouser (1964) compared three methods of teaching a programed unit on the laws of exponents to an accelerated track of 117 students from a single junior high school. The students were randomly assigned to three treatment groups, two discovery and an expository (rule-example) method. No verbal description of the discovery dimension was reported. By examining the learning programs, it was determined that the strategies were inductive in nature; sequences of examples were used to suggest a pattern from which the student was to generalize. The two discovery methods differed in that one method required the student to verbalize his discoveries whereas the other did not. The discovery programs were not completely programed; the first time a student had an opportunity to indicate whether or not he had discovered a particular rule, he did so by raising his hand. A proctor then checked his answer and, if correct, referred him to the next page. This procedure was followed so as not to present any clues to the learners, but may have created inhibitive effects for other learners.

The experiment was carried out during the regularly scheduled mathematics period. Two class periods were used for instruction. Because of snow, there was a one-day break in between.

Students were told that they were taking part in an experiment and that their performance would effect their mathematics grade. Failure to randomize the time factor, failure to control the Hawthorne effects, and providing external motivation all contaminate the treatments.

Four posttests (designed to measure manipulative ability, understanding, ability to transfer, and retention) were used to evaluate the relative effectiveness of the programed strategies. The transfer test was given five days after instruction while the retention test was given six weeks after instruction. The retention test contained 30 multiple choice items, most of which were nearly alike the questions asked in the experimental units to measure manipulative ability. A few questions were variations of the unit questions and checked understanding of the rules. The transfer test was a computation test. With the help of a table of values (viz., $2^1 = 2$, $2^2 = 4$, $2^3 = 8$, . . . , $2^{15} = 32{,}768$), students were to solve problems such as "$2^{14}/2^{10}$ = ___" and "What is the product of 16 and $2^8$?" There is some doubt as to whether this is a transfer test or a test of understanding (a minimal test of transfer). This author would call this type of transfer near transfer. The test was timed and the students were only given credit for those problems solved correctly by using the shortcut or rule. This scoring procedure seems undesirable for several reasons; the slow, methodological student is penalized, it encourages speed at the expense of accuracy, and the test is difficult to score. Concerning the scoring, Neuhouser (p. 25) stated," . . . This could be determined by the absence of the long computation, although usually there were other indications on the paper that the shortcut had been used." Many problems were instances or simple consequences of one of the four rules taught. The main effects were analyzed by using pairwise z- and t-tests, an inappropriate procedure to follow. No power levels were reported.

According to Neuhouser, the results indicated that the nonverbal discovery method was significantly superior to the expository method on posttests measuring retention, transfer, understanding, and on the total of posttests measuring manipulative ability, retention, transfer, and understanding. No other significant results were reported.

In summary, the experiment was poorly controlled and contained a questionable analysis of the data. The findings should be considered as tentative; they suggest the superiority of discovery teaching strategies, as judged by retention and near-transfer measures, compared with an expository strategy.

This author reanalyzed Neuhouser's data by using a one-way MANOVA design with equal sample sizes. No over-all significant differences were found at the .05 level of significance. To forestall any alternate conclusions, it was decided to check the homogeneity of covariance matrices assumption, a necessary assumption for the multivariate significance test. It was found that the matrices were heterogeneous, thus raising serious concern regarding the one-way multivariate test. No variance stabilizing transformations were tried.

Both of the studies by Sobel and Neuhouser contain problems in design and data analysis. As a result, no clear-cut evidence emerges to support either of the general teaching strategies for retention of mathematics at the junior high school level.

High School Level (grades 10 - 12). Three comparative studies investigated retention of mathematics at the high school level. The reported findings of these studies concerning retention are listed in Table 3. Two of these three studies yielded statistically significant results concerning retention at the high school level. These are the studies by Kersh (1962) and Hirsch (1972).

Kersh (1962), employing the same learning tasks used in his 1958 study, taught 90 high school geometry students by three different methods (directed learning, guided discovery, and rote learning) in an attempt to study the motivating power of guided-discovery methods. Prior to applying the three treatments, the students were told the two rules involved in the study and were given practice in their application. All students were taught until a criterion task of six successive applications of each rule

Table 3

Method Favored as Determined by the Significant Findings Reported by Experimenters for Retention of Mathematics

High School and College Levels

(≠ denotes College Studies)

|  | Discovery | Expository | Neither |
|---|---|---|---|
|  | Hirsch (1972) | Kersh (1962 | Kellogg (1956) |
|  | ≠Caruso (1966) |  |  |
|  | ≠Hanson (1967) |  |  |
| Total | 3 | 1 | 1 |
| Per cent (high school) | 33.3 | 33.3 | 33.3 |

was reached. The students were then randomly assigned to each of the three treatments. The directed-learning group was taught the rules and their explanation by programed learning with correct answers as feedback to the learner. Subjects in the guided-discovery group were required to discover the explanations and were taught tutorially using a form of Socratic questioning. The students in the rote-learning group were not instructed on the explanations of the rules; this group was included to act as a control for "meaningful learning." Kersh's description of the rote-learning group is not clear; this group could have learned by pure discovery.

After the learning period, a test of recall and transfer was given three days, two weeks, and six weeks later. For this purpose, each treatment group was subdivided into three subgroups of 10 students each, and each group took exactly one delayed retest. The test, similar to the one used in Kersh's 1958 study, contained two problems and a questionnaire. The number of subjects in each group who used the appropriate rule in an acceptable way on the test was used as the index of transfer; computational accuracy was not required. The criterion measure of retention was the number of subjects in each group who wrote an acceptable statement of each rule.

The data were analyzed by using a chi-square technique similar to ANOVA; the chi-square was broken down into the differences between teaching methods, the differences between test periods, and the differences attributed to interaction of treatments and time periods. No summary table of this technique was reported. It is, therefore, impossible to determine if his findings are in keeping with the data.

The results, according to Kersh, indicated that the rote-learning group was consistently superior in every respect to the other treatment groups. The guided-discovery group appeared superior to the directed-learning group on the three-days transfer and retention measures. Kersh suggests that retroactive inhibition accounted for the superiority of the rote-learning group:

> . . . The experimental efforts to inject meaning into the rules amounted to following their initial rote learning with a closely related and complex learning task; thus the rote learning group may have surpassed other groups simply because retention among the later was inhibited by the interpolated learning. (p. 69)

Kersh further suggests that " . . . The present results leave no doubt that there is a tendency for interest to accrue as a result of learning by discovery." (p. 70) This is open to debate. Interest may have accrued because of the novelty of the learning tasks.

Kersh confounded this study by not using a uniform mode of presentation across the three treatments. His findings, if valid, cannot be solely attributed to the three treatments. There is the possibility that the mode-of-presentation effects, alone, or in conjunction with the treatment effects, caused the noticeable differences in the dependent variables.

Hirsch (1972) compared the effectiveness of three modes of instruction on learning, transfer, and retention of certain subject matter concerning complex numbers. The three modes of instruction involved both teacher-presented and individualized, programed instruction. The two programed methods were expository in format with one of the methods employing a branching program. The teacher-presented method employed discovery-teaching strategies with emphasis on teacher-student development of definitions, mathematical principles, and generalizations through dialectical dialogue. Hirsch defined dialectical dialogue as

> . . . open, verbal sharing of ideas between teacher and students and between students themselves as they work together to arrive at definitions, mathematical principles and generalizations. (p. 61)

Commenting on the discovery strategy, Hirsch stated:

> . . . This approach, based on the logical structure of mathematics, involved deductive as well as inductive methods. The teacher's role was to lead the students through a predetermined sequence of questions and problems, subject to feedback from the students. . . . Class members were called upon to suggest definitions, to suggest properties, and to suggest methods of proof. (p. 62)

Six 55-minute class periods constituted the instructional period.

An attempt to control the Hawthorne and novelty effects was made. Schools and treatments were matched so that the subjects in each treatment had prior experience in a program which employed similar modes of instruction. Two hundred and thirteen students from six high schools were involved in the study. Classes of intact students were used at each school, with at least two classes assigned to each of the three treatments. By using the ITED Quantitative Thinking Test and the Cooperative Mathematics Test--Algebra II as pretests, an "eyeball" inspection of the means and standard deviations indicated that the six schools were matched. This is questionable since there was a range in means of five points on one measure and seven points on the other. The student score was used as the experimental unit for the study. This unit is questionable, particularly for the discovery method. Concerning this, Hirsch stated:

> . . . The use of the subject as the unit of analysis in the case of the guided-discovery treatment is debatable. However, not all students have the same insight during a discovery discussion and thus discoveries and methods of discovery differ from subject to subject. The foregoing observation and more importantly consistency in the analyses of the data suggested that the subject also be used as the unit of analysis in the case of the guided-discovery treatment. (p. 78)

This author cannot follow Hirsch's line of reasoning. Instead of fitting the data to a model, a model should be chosen to fit the data and design. Of course, in designing an experiment one keeps in mind both the data and the model to analyze the data.

The outcomes of instruction were measured by four measures: initial learning, vertical transfer, lateral transfer, and six-weeks retention. The retention measure had a reliability coefficient of .68, and the initial learning measure had a reliability coefficient of .75. No reliability coefficients were reported for the transfer measures. The retention test contained 25 multiple-choice items and measured concepts and generalizations developed in the lessons. The vertical-transfer test consisted of 11 problems which were more advanced than those used in the lessons, but required no more specific knowledge to solve. This test was designed to measure the student's ability to perform new tasks which had elements in common with the lessons, but which the basic procedure or method of solution should not be immediately apparent. The lateral-transfer test consisted of seven items to evaluate the student's ability to generalize to mathematical structures possessing similar structural characteristics. For each criterion measure, the data were pooled from the six schools into three grand treatment groups and analyzed by using the ANCOVA model. No justification was given for pooling the data.

The results of the study indicated that the guided-discovery groups were significantly (.01 level) superior to the other two groups on initial learning, vertical transfer, and lateral transfer. The guided-discovery group was found to be superior (.065 level) on the retention measure to either individualized instructional treatment group.

The use of intact classes and the student score as the experimental unit makes suspect any inferences drawn from the data. The results should only be considered as tentative.

In summary, neither the study by Kersh (1962) nor the study by Hirsch (1972) contain findings concerning retention that are not in some way suspect due to design or data analysis problems. In addition, the retention finding by Hirsch was significant at the .065 level, exceeding the experimental standard of .05 for developmental research studies.

College Level. Two comparative studies concerning retention were identified at the college level. These studies, by Caruso (1966) and Hanson (1967), both reported statistically significant findings concerning retention favoring the discovery teaching methods. These studies are listed in Table 3.

Caruso (1966) compared two methods (abstract and concrete) of teaching the theory of groups, rings, and fields to 186 college freshmen. The abstract approach presented the theory deductively by first giving rigorous definitions and then illustrating their properties and giving examples. Student participation under this method was kept at a minimum. The concrete approach presented the theory inductively by first presenting specific, concrete examples, becoming familiar with the theory by problem solving, and then introducing the student to the definitions and properties of the structures. In the concrete approach, the instructor encouraged student involvement, and solicited his aid in arriving at generalizations and conclusions.

The four-week experiment consisted of five abstract (experimental) and five concrete (control) classes taught by five instructors; each instructor taught an experimental and a control class in an attempt to control the teacher variable. In order to form comparable groups with respect to age, set, and achievement in mathematics, a limited number of subjects were changed from one class section to another (p. 105). No attempt was made to match or otherwise compare instructors. According to Caruso, instructors were not told which classes were experimental and which were control, but the experimenter would have had a hard time concealing this fact since he shared an office with three of the instructors. Very little control was placed on the teacher variable.

An achievement test (Test A) was given after the formal-learning period and was followed by a parallel form (Test B) of the achievement test, given nine weeks later as a retention test. Reliability coefficients of .74 for Test A and .505 for Test B were reported. A reliability coefficient of .505 for the retention test is low; 49 per cent of the variance in test scores is due to error variance.

The experimental unit for the study was the student score. After pooling the results into two groups (experimental and control), two t-tests were used to analyze the data. The experimenter reported no preliminary checks for homogeneity of means and variances for Tests A and B in order to justify pooling the data. Pooling the data increased the power of the significance tests. A significance level of .05 was set for the study, and a power level of .90, based on an N of 86, was reported. Caruso concluded that the experimental group was significantly (.06 level) better than the control group on Test A (the achievement test) and significantly (.01 level) superior to the control group on Test B (the retention test).

The use of the student score as the experimental unit is highly questionable in view of the fact that the treatment was group presented and allowed for student interaction. Thus, it is not likely that the student scores were independently distributed, a prerequisite for using the t-test. This, coupled with a low reliability coefficient for Test B, makes any interpretation of the data suspect. The reported findings are, at most, tentative.

Hanson (1967) compared discovery and expository techniques for teaching the constituent concepts involved in arithmetic sequences. The subject-matter content consisted of term, common difference, arithmetic means, general term, the determination of arithmetic means of two given numbers, and the summation of the terms.

The mode of presentation was by two individually programed units having identical formats but differing in presentation. In the discovery program, examples of each concept were presented from which the student had to induce the defining attributes of the concept. In the reception program, these same attributes were given to the student. The amount of overt activity in each program was equalized, thus indicating that differences, if they exist, are due to the manner in which the concepts are presented. Exercises followed the presentation of each concept or group of concepts.

The students exposed to the discovery treatment were divided into two groups for the purpose of investigating what effects, if any, verbalization has on learning, transfer, and retention. The students in one group were required to verbalize (define or explain) each concept immediately after discovering it. The students in the other group were not required to verbalize their discoveries.

The experiment was conducted twice, once with eighth grade students and once with college students. One hundred and four eighth-grade students in an advanced track from a single school were randomly assigned to the three treatments, and 107 college students (elementary education or nonscience majors) who had weak backgrounds in mathematics were used in the replication of the eighth grade study. No rigid time requirements were imposed on the student in either study.

Initial learning, transfer, and two-week retention measures were used to evaluate the effectiveness of the programs. The retention test was an equivalent form of the initial-learning test. No reliability data were reported for any of the criterion measures. The transfer test consisted of four parts: generalization, reversibility of operations, applications to real life situations, and analogous learning situations.

The data were analyzed by using ANOVA and ANCOVA models. For each study, nine Cochran C-tests at the .05 level were used to check homogeneity of variance assumptions; two of these 18 tests yielded significant differences, thus violating ANOVA assumptions. The heterogeneity of variance in two cases raises questions concerning the validity of the ANOVA model to analyze the data. Eighteen F-tests were employed for each study to test for main effects, making a total of 36 F-tests of significance conducted during the study, and raising serious questions concerning the size of the systematic experimental error rate. By chance factors alone (assuming independence of the tests), two of these tests could have resulted in rejecting the null hypothesis. There was no control group incorporated into the experimental design, and no power level was reported for any of the tests used.

Hanson reported that the college discovery group mean scores were significantly (.01-.05 level) higher than those of the college reception group on each total criterion measure. Only the reversibility transfer subtest yielded significant results, and the discovery group was significantly

(.01 level) superior to the reception group on this measure. There were no significant results found for the eighth grade students. There were no significant differences between the discovery-verbal and the discovery-nonverbal groups on any of the measures. An analysis of the test items led to the conclusion that inductive-discovery learning enhances the formulation of operational concepts to a greater extent than it does classification concepts. The eighth-grade students performed almost as well as the college students on exercises involving classification concepts. Operational concepts were too difficult for eighth grade students regardless of the way in which they were learned.

With no reliability data available, and with a questionable analysis of the data making any interpretation of the data suspect, Hanson's findings should only be considered as suggestive evidence.

Neither of these studies offer unquestionable support for the discovery method on the retention dimension. Taking Hanson's study at face value, one could not determine whether his discovery treatment was better than no treatment on the retention measure since a control group was not included in the study. It is conceivable that a control group might have been as effective as either treatment, especially at the eighth-grade level where the content was considered too difficult for the students.

Transfer

Twenty-eight comparative research studies investigated transfer of mathematical training. Only one study (Worthen, 1965) investigated the effects of negative transfer. Many different criterion measures were used to assess transfer. Among these are the ratio of the number of correct applications of the rule to the number of correct answers (Hendrix, 1947), methods used in solving problems (Kersh, 1958), a comparison of each group's own performance at different times during the study (Swenson, 1949), time to solve problems (Gagne and Brown, 1961), the number of hints required for criterion (Gagne and Brown, 1961), a weighted-time score (Eldredge, 1965 and Meconi, 1967), and raw scores from transfer tests. A majority of the studies compared raw scores from transfer tests.

Transfer measures were employed immediately after instruction and at points up to six weeks after instruction (Kersh, 1958). Hendrix administered her transfer test approximately two weeks following instruction.

Elementary Level (Grades K - 6). Twelve comparative studies investigated transfer of mathematical training at the elementary level. Eight of these studies reported significant findings concerning transfer of mathematical training, in every case favoring discovery methods. Table 4 classifies the transfer findings at the elementary level. The studies by Fullerton (1955), McConnell (1934), Norman (1955), Thiele (1938), Worthen (1965), and Winch (1913) were reviewed under the section on retention of mathematics at the elementary level.

Anderson (1949), in a study lasting approximately seven months, compared two techniques, drill and meaning, for teaching arithmetic to approximately 389 fourth-grade students in 18 classes from 18 public schools. By labeling the procedures drill and meaning, the experimenter biased the experiment before it got underway. Anderson did not give a clear characterization of each teaching procedure. The drill method presented units of arithmetic in an unconnected fashion; the student was to master these discrete units through formal repetition. The meaning method presented arithmetic as a closely knit system of ideas, principles, and processes learned from the beginning in a meaningful fashion with students encouraged to discover relationships.

## Table 4

### Method Favored as Determined by the Significant Findings Reported by Experimenters for Transfer of Mathematical Training

### Elementary Level

### (Kindergarten to Sixth Grade)

| Discovery | Expository | Neither |
|---|---|---|
| Anderson (1949) | | Bassler *et al.* (1971) |
| Fullerton (1955) | | Peters (1970) |
| McConnell (1934) | | Scott (1970) |
| Norman (1955) | | Twelkner (1965) |
| Swenson (1949) | | |
| Thiele (1938) | | |
| Winch (1913) | | |
| Worthen (1965) | | |
| Total 8 | | 4 |
| Per cent* 67 | | 33 |

*Based on a total of 12 studies.

There was very little control of the teacher variable. Teachers were assigned to treatments on the basis of interviews and a measure of attitudes. Concerning instructional procedures, Anderson stated:

> . . . Day-by-day procedures for each method were not prescribed, nor were the teachers held to parallel courses concerning objectives, content, amount of time spent on instruction and on drill, number of repetitions of the number facts, and the textbooks and other instructional material. Broad objectives were set for each method and relatively free reign was given teachers, within the limits of the basic theory of learning and its application to arithmetic instruction, to achieve these objectives. . . . (p. 44)

The dependent variables were test scores on tests of computational skills, problem solving, understanding of social concepts in arithmetic and vocabulary, and of mathematical thinking (transfer).

Following a confusing and questionable analysis of the data, Anderson concluded, among other things, that on the tests of mathematical thinking, the difference between the high-ability groups significantly (.01 level) favored the meaning method, whereas the difference between the low-ability groups significantly (.01 level) favored the drill method.

Due to the lack of control of the independent variables, it is impossible to draw any useful conclusions from

Anderson's study. Furthermore, from his report of the experiment, it would be difficult, if not impossible, to replicate the study.

Swenson (1949), in a 20-week study, taught 100 addition facts to 332 second-grade students from 14 classes. Three different methods were employed: generalization, drill, and drill-plus. The generalization method encouraged students to build up interrelationships among addition facts which were presented in groups determined by some unifying principle. Students were not given the generalization, but were encouraged to formulate their own generalizations. The drill method presented the addition facts as "facts to be learned" in a random order. The drill-plus method was to duplicate common practice, the drill method with certain concessions to the ideas of concrete meaning and organization. Students were allowed to use concrete models to verify the number combinations, after which drill procedures followed. Further, number combinations yielding the same answer were grouped together, but children were discouraged from forming generalizations; this practice placed a strong experimental bias in favor of the generalization method.

There was little control of the teacher variable, and the classes were assigned at random to the different methods of instruction. Teachers had no choice of method.

Tests were administered at eight different times during the experiment. A test consisting of 100 addition facts was administered by using flash cards at five different times during the study: as a pretest, after instruction on an original set of facts (O), after learning an interpolated set of facts (I), after two-and-a-half weeks of vacation, and after instruction on a final set of facts (F). After instruction, three transfer tests were given: a 100-item subtraction facts test, administered by flash cards; a 100-item decade addition test (one-digit number plus a two-digit number) was administered in mimeograph form; and a 40-item test consisting of a variety of addition problems. The transfer tests not employing flash cards were power tests. It was reported that each test had a reliability coefficient of at least .95. Since a control group was not included in the study, transfer was measured by comparing each group's own performance at different times during the experiment. Concerning transfer, Swenson stated:

> Transfer in this study refers to gains in the performance of any group in their knowledge of a certain set of facts during instruction on another

> set. It may also refer to the mean gain in performance during the vacation period if a gain appears for that interval. (p. 10)

Further, Swenson makes a finer distinction of transfer and states:

> Transfer is used here to refer to positive mean gains in knowledge of a certain set of facts during a period when they have not been studied directly. Retroactive inhibition here refers to negative mean changes on knowledge of a certain set of previously learned facts. The writer is aware that negative transfer effects may be concealed within the present results. The result sometimes referred to as "retroactive facilitation" shows up here under the term "transfer to previously taught facts." (p. 25)

The data were analyzed by using gain scores and the ANCOVA model. Pairwise t-tests were used for post hoc comparisons following a significant F-ratio for main effects. Swenson included no statistical data in her report, so one cannot determine whether her findings are in keeping with the data.

Swenson concluded, among other things, that (1) there was no significant difference in retention between the drill-plus and generalization groups when retention was considered for the entire remainder of the study following initial instruction, (2) the transfer within the addition facts favored the generalization method on most of the tests, and (3) transfer to untaught subtraction facts showed that the generalization group was significantly superior to the drill method, and that the drill method was significantly superior to the drill-plus method.

Swenson's account of her findings is complicated and confusing. Without a summary of the statistical data, it is difficult to accurately determine just what took place.

In summary, after analyzing the experimental designs and data analysis for the eight studies at the elementary level claiming statistically significant findings for transfer favoring the discovery methods, one cannot rule out the possibility that each of these findings resulted from confounding and contaminating of variables or faulty data analyses.

Junior High Level (Grades 7 - 9). Eight comparative studies were found that investigated transfer of mathematical training at the junior high school level. These studies are listed in Table 5. As can be seen from the table, if a transfer finding was reported as statistically significant, it, in four of five cases, favored the discovery methods. The study by Gagne and Brown (1961) involved only male students from both the ninth and tenth grades, and the study by Wolfe (1963) involved students from both the junior and senior high school levels. The study by Neuhouser (1964)

## Table 5

## Method Favored as Determined by the Significant Findings Reported by Experimenters for Transfer of Mathematical Training

### Junior High School

### (Grades 7 - 9)

| | Discovery | Expository | Neither |
|---|---|---|---|
| | Eldredge (1965) | Michael (1949) | Brown (1969) |
| | ≠Gagne and Brown (1961) | | Meconi (1967) |
| | Kuhfittig (1972) | | ##Wolfe (1963) |
| | Neuhouser (1964) | | |
| Total | 4 | 1 | 3 |
| Per cent* | 50 | 13 | 37 |

≠Ninth and tenth grades.

##Junior and senior high school Ss.

*Based on a total of 8 studies.

was reviewed in the previous section on retention of junior high school mathematics.

Michael (1949) compared an inductive-discovery method and a deductive-expository method for teaching positive and negative numbers, the fundamental operations with them, and the solution of simple equations to fifteen, intact classes of ninth-grade algebra students. The discovery method

> . . . emphasized the use of exercises in thinking, with the exercises built around familiar situations involving time, money, directions, temperature, and others of the type commonly used in textbooks in algebra. Through the use of these exercises, the pupil was expected to discover and understand the fundamental principles and relationships to be learned. The use of numerous practice exercises to bring about efficiency in the operations was supposed to follow the discovery and understanding brought about inductively by the learning exercises. While pupils undoubtedly came to generalize for themselves individually at various times during the experimental period, no statement of the rules of operation was made by teachers or pupils in teaching, reteaching, or pupil discussion. (p. 83)

The expository method

> . . . emphasized the use of authoritative statements of the rules of operation combined with extensive practice or drill. No attempt was made, before practice with the respective processes was begun, to explain why the rules operated to give the correct results. Through the process of working with the rules in many exercises the pupil was expected to gain operative efficiency and to acquire understanding of the principles and relationships in the area under consideration. (p. 83)

The experimental period consisted of approximately 45 class periods.

Computation, generalization, and attitude (toward algebra and toward mathematics) measures were used to compare the relative effectiveness of the two methods of instruction.

The computation test consisted of six subtests involving the fundamental operations, and the generalization test consisted of seven subtests. The generalization test included applications involving substitution of equivalent expressions, interpretations of signed numbers, generalized expressions involving signed numbers, and interpreting functional relationships. The computation and generalization tests each included 100 items, while the attitude test contained 20 items. All tests were used as both pretests and post-tests. Corrected, split-half reliability coefficients for the computation and generalization tests at both administrations ranged from .94 to .96. The reliability of the attitudes test was estimated from a test-retest situation; the correlation of the two sets of scores was .91.

According to Michael, a manual was prepared for the participating teachers outlining the methods and providing specific instructions for teaching each section under each method. This was done in an attempt to control the teacher variable.

ANCOVA was used to analyze the data. Part of the analysis involved checking the differential effect of methods at different ability levels. Concerning this, Michael stated:

> Three equal levels of ability were established on the basis of total intelligence test scores, and the groups obtained by sorting the punch cards for each method on this basis were studied. These analyses were limited to computation and generalization test scores. (p. 86)

No justification was reported for this questionable pooling of data. Further, he indicated that disproportionate subclass frequencies resulted (p. 86); no mention was made of establishing the necessary assumptions for the ANCOVA.

Michael reported that the discovery method was significantly (.05 level) superior to the expository method on the multiplication subtest of the computation test, the expository group was significantly (.05 level) superior to the discovery group on the generalization test, and the discovery groups had a significantly (.01 level) better attitude toward algebra than did the expository groups. No other statistically significant differences were reported.

Michael's report of the statistical analysis is not clear. From his abbreviated summary tables, one cannot determine whether his findings are in keeping with the data. He does not indicate an alpha level for any of his tests, nor does he indicate the experimental unit for any of his tests, although it appears that the student-score was used. Further, from his report, one does not know whether the

ANCOVA model was appropriate for analyzing the data; no homogeneity assumptions were reportably satisfied.

The use of ANCOVA and intact cleasses does not eliminate the necessity of using random samples. Concerning this, Peckham _et al_. (1969) state:

> . . . While ANCOVA may increase the sensitivity of comparisons in many instances, Lord (1967) and others have shown that in the absence of the tenability of the assumption stated earlier (the use of random samples), the analysis of covariance can result in misleading conclusions. (p. 347)

There is some question as to the reliability of Michael's subtests. He reported only reliability estimates for his total-tests. It is generally the case that subtests have lower reliability estimates than the total-test. He indicated (p. 85) that a test-retest reliability coefficient of .91 was computed for the attitudes test. This would seem to imply that the measure is a fairly stable predictor of attitudes, although Michael reported (p. 86) that the discovery method produced significant changes in the attitude of the students toward algebra. Thus, one would suspect that the attitude-toward-algebra subtest was not a reliable measure.

From Michael's description of his generalization test, it is not clear whether this test should be classified as a transfer test. The test appears to be testing learned knowledge at the applications level. If this is the case, the term generalization test is misleading. Concerning the terms generalization and transfer, Edwards and Scannell (1968) stated:

> Generalization and transfer are sometimes used as synonymous terms. . . . Generalization is defined as a "general" process which occurs in human behavior. The individual reasons from an event that "this follows . . ." and acts on that basis . . . Transfer of training is defined as the adaptation of specific reactions to situations other than those to which they were originally specific. As such, transfer becomes a special case of accurate generalization. (p. 356)

No conclusive, supportive evidence can be drawn from this study.

Gagne and Brown (1961) taught principles pertaining to number series to three groups of ninth- and tenth-grade students. Three programed units (rule and example, guided discovery, and discovery), all consisting of a common introductory program, were designed to facilitate the learning of the concepts used in deriving formulas for the sums of terms of unfamiliar number series. Eleven boys were randomly assigned to each treatment group, for a total of 33 subjects. The introductory program contained 89 items and was intended to establish the learning of basic definitional concepts that would be used in later problem-solving. For the training program, the subjects in the rule-example group were first given the correct formula followed by several examples for the student to work. The students in the discovery group were required to discover the rules for summing the number series. Hints were provided to direct the discovery students to work at deriving the rules for summing several different series. Hints given to the guided-discovery students were more specific in nature; they were taught how to discover the rules.

On the second day of the two-day training program, each student repeated the first day's program in order to maximize the probability of learning the lesson.

Immediately after the learning program, a transfer test, consisting of four new series, of discovery skill was given. The test measured the learner's ability to discover new rules from different problem series. These problems were given one at a time, and a time interval of 10 minutes was allowed for completing each problem. Hints were available, on separate cards, if the students felt the need for help. For each problem, students were instructed to show their answers to the experimenter; he would tell them whether their answers were correct or not and whether or not to continue working on the problem.

The treatments were evaluated by using three criterion measures: time to solve the problems, the number of hints required to discover the rules, and a weighted-time score. No reliability data were reported for either the time or number of hints measure. A split-half reliability coefficient of .72 was reported for the weighted-time score.

The data were analyzed by using three ANOVA designs, followed by pairwise t-tests for significant F's. The experimenters concluded that the guided discovery and discovery groups were significantly (.01 level) superior to the rule and example group on all measures. For the time scores, the guided-discovery group was significantly (.02 level) superior to the discovery group.

Gagne and Brown's use of pairwise t-test comparisons is not an appropriate model to follow after a significant F-ratio since they lead to spurious probability statements. Scheffe's test for pairwise differences would have been more appropriate.

The students in the guided-discovery treatment had an advantage over the students experiencing the other two treatments for solving the novel transfer tasks. They were presented with pointed cues which, if followed, instructed them how to sum the number series. It is no wonder that this experimental group excelled on the transfer tasks; these students got more practice on summing number series. You might also say that their learning was more efficient in terms of the time to learn.

This appears to be a well-designed experiment. Concerning the analysis of the data, one cannot be sure of the true alpha level for each main hypothesis; it is somewhere between .05 and .20. This study offers suggestive evidence that particular, programed-discovery strategies are better for horizontal transfer of certain rule learning to the learning of new rules than a particular programed expository strategy.

Eldredge (1965) compared two methods of programed instruction, guided discovery and rule-example, for retention and transfer of number series tasks. The two learning programs were revisions of those used by Gagne and Brown (1961). Each program contained a common introductory program, and students were required to go through the treatment programs twice. Learning materials were printed on 3"x5" cards with the correct responses to the frames printed on the back of the cards. The guided-discovery method, according to Eldredge,

> . . . incorporated a sequencing of items in which the subject was given specific cues to aid in discovering a rule and formula for summing number series before actually being shown a statement of the rule. The cues were in the form of Socratic questions such as "Can you get 6 from 3 by adding, subtracting, multiplying, or dividing?" (p. 7)

Students learning under each technique were given equal amounts of practice in solving problems and were required to solve the same problems. All students were exposed to the treatments for three class periods. There was a deliberate attempt to teach a searching behavior on the part of the guided-discovery program.

Ninty-six students from two intact, ninth-grade algebra classes, one from each of two schools, were randomly assigned to the two treatments. All students were considered above average in ability.

Students in both treatments were told that they were taking part in an experiment. This failure to control the Hawthorne effect possibly contaminated the treatments.

To test the differences between treatments, tests of immediate transfer, delayed transfer, and delayed retention were administered. The initial-transfer test consisted of four problems which required the student to find formulas for summing number series which were different from the four series involved in the learning task. Guided-discovery students were told to raise their hands when finished with a problem; a proctor then checked their work and told them if their answer was correct or not. The initial-transfer test had an alpha reliability coefficient of .56 (The alpha coefficient is the mean of all possible split-half coefficients resulting from the different splittings of the test.). Following the initial-transfer test, the subjects from one school were given an opportunity to explain orally the processes they used to solve the problems. The delayed-transfer test was given four weeks after instruction and had an alpha reliability coefficient of .55. Both transfer measures were scored using weighted, time-hint scores, high scores indicating poor performance.

The delayed-retention test consisted of the same four series used in the learning task with no hints provided. This test was given four weeks after the learning program. Scores on the retention test consisted of the amount of time needed to solve a problem plus a one-minute penalty for no solution and a two-minute penalty if the solution was correct as to the process but wrong as to terminology. The time limit for each problem was five minutes, and the maximum score obtainable was 32, eight points for each problem.

Following a questionable procedure, the data were analyzed by using the ANOVA and ANCOVA models. The time spent in the learning programs was used as a covariate in evaluating each criterion measure; this variable was clearly influenced by the treatments and represents a questionable move. The guided-discovery students were found to be superior to the rule-example students on both transfer measures. In an attempt to disprove Kersh's hypothesis that guided-discovery students are more highly motivated and continue to work on the tasks after instruction, Eldredge used the ANCOVA model on the delayed-transfer scores with the initial-transfer scores as the covariate, also a questionable procedure. Eldredge stated:

> . . . When the covariance design controlled for differences between treatments on the initial transfer test, the difference over time disappeared. That transfer favoring the guided discovery treatment is not disputed. However, the delayed transfer differences cannot be explained in terms of enhancement of motivation as hypothesized by Kersh. . . . (p. 50)

No significant differences were reported for the retention measure, but students required to verbalize the principles and processes involved in their discoveries improved their scores on the retention measure significantly more than subjects not required to orally verbalize their findings.

This experiment is well-designed, but did not control for the Hawthorne and novelty effects. This, and a questionable use of the ANCOVA model, makes suspect any inferences drawn from the data. Furthermore, no power levels were reported for any of the significance tests.

Since correct responses were supplied for each treatment frame, it is impossible to ascertain whether the students under the guided-discovery treatment discovered anything. This experiment may have compared two expository methods of programed instruction.

Kuhfittig (1972) investigated the relative effectiveness of guided-discovery learning, compared to expository learning, and the relative merits of using concrete learning-materials, as opposed to teaching abstractly, when teaching seventh-grade students a unit on converting American currency to old-English currency and vice versa. Forty seventh-grade students from an elementary school were involved in the study. A pretest was administered to identify and remove students who had an acquaintance with English currency. To insure the existence of two distinct ability levels, only those students whose mean score on the arithmetic reasoning and computational skill subtests of the MAT was one standard deviation from the population mean were included in the sample.

The teaching and learning-aids dimensions resulted in four groups, each with five students from each ability level. Students were randomly assigned to each group.

The discovery groups were taught by a carefully structured sequence of questions, whereas the expository groups were taught by careful explanations of individual steps. Kuhfittig taught all groups in order to insure uniformity of instructional procedures. All instructions were presented to the groups by means of an overhead projector and supplemented by oral instructions. A review of audio recordings of the lessons revealed that the experimenter was faithful to each treatment. Each treatment was administered for two class periods on consecutive school days.

On the learning-aids dimension, the concrete groups were allowed to manipulate coin models of the currency, while only verbal references were used in the abstract groups. In addition, members of the concrete classes were required to give responses in terms of their models.

After the learning session, a posttest consisting of achievement, horizontal transfer, and vertical transfer subtests was administered. Four weeks later, the same posttest was readministered as a retention measure. The achievement subtest contained 29 items that were similar to those used in the lesson. The horizontal-transfer subtest contained 15 items, 12 of which were similar to the items on the achievement test. Three new terms and definitions were involved in the remaining three items. The vertical-transfer subtest contained 10 items designed to determine whether the students could convert British to German currency. Reliability estimates for the achievement and horizontal- and vertical-transfer subtests were .82, .84, and .55, respectively.

The data were analyzed by using three 2x2x2x2, ANOVA designs with repeated measures on the last factor and with five subjects per cell. Each design considered one subtest and its retention-posttest counterpart. For each repeated-measures design, the scores on the two posttests were also treated as separate 2x2x2, ANOVA designs. In total. 15 ANOVA designs were employed to analyze the data, not including the homogeneity of variance assumptions. No significant results were found concerning the treatment effect. Kuhfittig reported, among other things, that the guided-discovery groups using concrete learning aids had higher mean scores on the transfer portions of the posttest than did the expository groups.

Several questionable procedures were followed in Kuhfittig's study. First of all, students were taught in four separate groups, and the data were analyzed by using eight separate cells of five subjects each. Secondly, his analyses of the data did not try to minimize the experimental error rate. Further, his use of a repeated-measures design possibly resulted in experimental bias, this being the result of test carry-over effects, between treatment activities, and the possible interaction between carry-over effects and between-treatment activities. No reported attempt was made to determine whether the 2x2 population covariance matrices were equal and of compound symmetry. Winer (1962) states:

> The model under which the usual F tests in a repeated measures factorial experiment are valid not only assumes that the matrix of covariance within each of the populations is $\Sigma$ (the same for each of the populations) but also that $\Sigma$ has the following form. . . . each entry on the main diagonal is equal to $\sigma^2$, and each entry off the main diagonal is equal to $\rho\sigma^2$. (pp. 370-71)

Lastly, since the instruction was presented in group format, using the student score as the experimental unit is also debatable.

Kuhfittig's questionable data analysis precludes any meaningful interpretations concerning transfer to be drawn from his study.

In summary, only the study by Gagne and Brown (1961) offers support for the superiority of certain discovery methods, over an expository method, for facilitating transfer of certain rule-learning to the learning of certain, new rules at the junior high school level. Even if Gagne and Brown's (1961) transfer findings were found to be significant at the .05 level, which is doubtful, the findings have limited generalizability--holding only for male students similar to those in the sample and students using materials similar to those used in the study in programed format.

High School Level (Grades 10 - 12). Only two studies (Hirsch, 1972 and Wolfe, 1963) investigated transfer of mathematical training at the high school level. The study by Hersch (1972) reported a statistically significant finding favoring a discovery method, whereas the study by Wolfe (1963) yielded no statistically significant results concerning transfer. Hirsch's study was reviewed in the previous section on retention at the high school level. This study displayed problems in experimental design and analysis of the data.

College Level. Seven studies were found that investigated transfer of mathematical training at the college level. These studies are listed in Table 6. Three of these studies reported statistically significant findings concerning transfer, all in favor of the discovery methods. The study by Hanson (1967) was reviewed in the previous section on retention of mathematics at the college level.

Table 6

Method Favored as Determined by the Significant Findings Reported by Experimenters for Transfer of Mathematical Training

High School and College Levels

(≠ denotes College Studies)

| | Discovery | Expository | Neither |
|---|---|---|---|
| | ≠Hanson (1967) | | ≠Craig (1956) |
| | ≠Hendrix (1947) | | ≠Gaston and Kolb (1973) |
| | Hirsch (1972) | | ≠Krumboltz and Yabroff (1965) |
| | ≠Kersh (1958) | | Wolfe (1963) |
| | | | ≠Woodward (1966) |
| Total | 1<br>≠ 3 | | 1<br>≠ 4 |
| Per cent | 50<br>≠43 | | 50<br>≠57 |

Hendrix (1947), in three separate experiments, studied the effects of verbalizing a discovered generalization on transfer power. Two studies used college students and one study employed eleventh and twelfth-grade high school students. In each study, students were assigned to three treatment groups, two discovery and one expository. A control group of six students was included in the high school study. The learning task was to learn the formula for the sum of the first n odd positive integers. Upon discovering the principle, one discovery group was asked to verbalize the discovery, while the other was not. The expository group had the principle told to them.

A transfer test was given approximately two weeks after instruction. Hendrix does not describe the transfer test, other than it contained 10 items, nor does she report any validity or reliability information concerning the test. A criterion used in evaluating the results of the transfer test was the ratio of the number of correct applications of the rule to the number of correct answers obtained by each student. It was found that three of the six control group members in the high school study acquired the generalization on an unverbalized level through counting and adding in the earlier test questions. Consequently, the last four items of the transfer test were omitted from the analysis of the data.

The results of the three experiments were pooled to obtain a sample of 40 scores. No justification for this pooling of the data was reported.

The analysis of the data included pairwise t-tests. The fourteen subjects taught by the expository method accounted for 47 applications out of 68 correct answers, a ratio of .69. Students taught by the subverbal awareness method accounted for 57 applications out of 65 correct answers, a ratio of .88, and the verbal discovery group accounted for 49 applications out of 65 correct answers, a ratio of .75. Standard deviations and the significance levels employed in the studies were not reported. Hendrix concluded that the verbal-discovery group was slightly superior (.33 level) to the non-verbal-discovery group on the transfer measure, the non-verbal-discovery group was superior (.12 level) to the expository group on the transfer measure, and the verbal-discovery group had higher transfer-effects than the expository group.

Hendrix's written account of her experiments, including the analyses of the data, is confusing and unclear on some points. It is not clear whether the students in her studies were randomly assigned to treatments or not. Further, she presents evidence to indicate that her groups were not matched on certain variables.

> In the first run of the experiment, differences in achievement marks and reading-test scores rated Group 2 (taught by Method II) slightly above Group 3 (taught by Method III) but Group 1 (taught by Method I) was above Groups 2 and 3 combined. (p. 198)

Any inferences concerning the superiority of the non-verbal-discovery method drawn from this study are suspect. It is not known why much research literature cites this study as supporting the superiority of the non-verbal-awareness method.

Kersh (1958) compared six treatments for teaching college students two addition rules. One rule determined the sum of the first n positive odd integers, and the other determined the sum of an arithmetic progression of positive integers. The six treatments resulted from crossing two methods of representing problems and three methods of teaching. The problems were either presented in an iconic form involving *X*'s, called the *X-form*, or the conventional, Hindu-Arabic form, called the *A-form*. The X-form revealed or suggested certain geometrical relationships.

A total of 60 college student volunteers from two sections of Educational Psychology taught by the experimenter formed the target population for the study. By using a table of random numbers, eight students were assigned to each of the six groups. The six groups were reported to have been judged equivalent in terms of age, sex, grade level, and scholastic aptitude.

Instruction was individualized and a time period of 60 to 90 minutes was scheduled for each student. Many students, particularly those in the no-help groups, were unable to discover the rules in the allotted time period. Six different problems, three for each rule, were used in the learning period, and they were essentially the same for all six groups. Each student was asked to "think aloud" during the learning period, and voice recordings were made.

Following the learning period, each student was given 20 problems for the purpose of detecting differences among the students in their achievement, if any. The 20 problems consisted of five odd-number-rule problems in A-form, five odd-number-rule problems in X-form, five progression-rule problems in A-form, and five progression-rule problems in X-form. The ten problems of each type were actually five problems presented once in the A-form and once in the

X-form. The odd-number-rule problems preceded the progression-rule problems. Students were first presented the problems in the form used in their treatment.

A retest was administered to all students four to six weeks after the first test. The retest contained two problems, which could easily be solved by using the two rules, and a questionnaire on their process of thinking. The datum of primary interest was the method used in solving the problems.

No validity or reliability information was reported for either test. All learning and test problems were reproduced on separate pieces of paper or cardboard so that they could be presented one at a time.

The retest data were analyzed by using eight chi-square tests of independence. A 2x3, methods (use of the correct rule or not) by teaching treatments (no-help, direct-reference, and rule-given), chi-square test for problem 1 (odd-number rule) on the retest was significant at the .05 level. This, according to Kersh, indicated that there was a significant relationship between treatments and methods for this problem (the progression-rule problem test was not significant). By inspecting the data, it is seen that the no-help groups used the correct rule more often than the other groups. There was also a significant (.05 level) relationship between correspondence of methods used on the retest with those learned during the learning period for the three teaching treatments. The no-help and direct-reference groups used the same or other methods more frequently than the rule-given groups. The other six chi-square tests yielded no significant results.

It was noted by Kersh that although 13 students in the no-help groups failed to learn an acceptable rule for the progression-rule problems during the learning period, there were only four who added on the retest, and 10 who used acceptable methods. Kersh attributed this change of behavior to the no-help group being motivated to continue practicing the task after the learning period. Kersh stated:

> . . . The results of this experiment suggest that when the learner is forced to rely on his own cognitive capacities, it is more likely that he will become motivated to continue the learning process or to continue practicing the task after the learning period. (p. 292)

No conclusive evidence can be drawn from this experiment concerning the most efficacious teaching method for facilitating transfer.

There are two tentative explanations for the change in performance of the no-help groups on the retest. First, the Ovsiankina effect (the resumption of incomplete tasks) may have been working. This is particularly plausible since the learning tasks involved novel problems. There is also the possibility that the no-help students communicated with other students involved in the experiment. This is also plausible since all students were members of the experimenter's two Educational Psychology classes. Kersh did not report whether he instructed the students that a retest would take place. If they were so instructed, this could have had a motivational effect on the students.

In summary, no unquestionable research support exists to substantiate a superior approach, either discovery or expository, for teaching college mathematics in terms of transfer effects.

## Science

This section will be devoted to analyzing and summarizing the comparative research studies investigating retention or transfer in the area of science. The studies will not be grouped by grade level, largely because of the lack of reported significant findings.

The discovery studies in the area of science take on an added dimension that most often is absent in discovery studies in mathematics. The added dimension is the use of the laboratory in instruction. As a result, there are essentially four teaching patterns that can result in comparative studies. These are: discovery lecture-expository laboratory, discovery lecture-discovery laboratory, expository lecture-discovery laboratory, and expository lecture-expository laboratory.

Various degrees of guidance may be provided for the laboratory session. These include giving, or not giving the problem; giving, or not giving the ways and means; and giving, or not giving the answers or results.

Frequently, the expressions open-ended experiments, discovery approach, free experimentation, and inquiry lessons are used interchangeably by some people. According to Lennek (1967), open-ended experiments are

> . . . those types of laboratory exercises for which the answers are not known nor are they to be found in textbooks or other reference sources. They require investigation on the part of the student and often times the results of these experiments lead to as many stimulating questions as satisfying answers. (p. 12)

Lansdown and Dietz (1965) state that in free experimentation

> . . . The student is faced with structured materials and told to "see what you can find out." The structure of the materials here leads the experimenter to discover that some things sink, some float, and some do both . . . (p. 211)

The discovery approach prevails upon the teacher for the source and direction of classroom transactions, as in the Socratic Method. In the inquiry-training approach (Suchman, 1961) to the teaching of science, the students are generally in control of these classroom transactions. After being shown a brief film strip of an intrinsically anomalous event, the students conduct an inquiry by interrogating the teacher in a format much like "Twenty Questions."

### Retention

As with the mathematical studies, only those comparative research studies assessing achievement initially following instruction and again at some later date (as a delayed achievement measure) will be considered for review.

Only eight comparative research studies were found that investigated retention of science learning. These studies are listed in Table 7. Only the study by Boeck (1951) reported significant results concerning retention.

Boeck (1951) contrasted an inductive-deductive method with a deductive-descriptive approach to instruction in high school chemistry. The inductive-deductive approach differed from the deductive-descriptive approach essentially in the laboratory phases. The inductive-deductive laboratory approach was student centered and progressed from particular

Table 7

Significant Findings Reported by Experimenters for Retention of Science Learning

| Study | Year | Level | Method Favored | Subject Area |
|---|---|---|---|---|
| Boeck | 1951 | H S | Discovery | Chem |
| Gentry | 1965 | Jr H | Neither | Gen Sci |
| Brudzynski | 1966 | Elem | Neither | Phy Sci |
| Lennek | 1967 | Jr H | Neither | Gen Sci |
| Brenner | 1968 | Coll | Neither | Phy Sci |
| Dennison | 1969 | Jr H | Neither | Gen Sci |
| Tanner | 1969 | Jr H | Neither | Phy Sci |
| Zubulake | 1970 | Jr H | Neither | Gen Sci |

instances to generalizations. Pupils were encouraged to follow scientific procedures. In the class periods following the laboratory work, discovered generalizations were applied to numerous related problems.

In the deductive-descriptive approach, the laboratory exercises were taken from a representative, published laboratory manual. The experiments were carried out after the general principles involved had been presented and discussed. No provisions were made for student's planning of experiments, and little or no opportunity was provided for the solution of real problems under laboratory conditions. It is not clear, from Boeck's report, just how the two methods differed; it would be difficult to replicate this study.

Forty-seven students were randomly selected from the chemistry enrollment at a single high school and assigned to the two treatments. Seven classes were selected at random from the other schools in the state having approximately the same enrollment as the experimental school to take part in the evaluation involving some one of the desired objectives; only the deductive-descriptive treatment was used at these schools.

Achievement-test scores were obtained at the end of the nine-month school term to measure the attainment of factual materials and principles, the ability to apply principles, and the ability to use the scientific method with an accompanying scientific attitude. Retention scores were obtained for those students who had not been graduated from high school four months after the end of the school term; Boeck does not indicate the number of such students. The achievement tests were reported to have a median reliability coefficient of about .80.

The data were analyzed using covariance analysis with I.Q. and pretest scores as covariates. Individual student scores were used as the unit of experimental analysis, a questionable procedure. No data were reported for the retention measure. Boeck concluded that the groups taught by the inductive-deductive method were found to be superior in demonstrating resourcefulness in the laboratory, applying the scientific method, applying generalizations, and mastery and retention of principles and facts. Concerning retention, one is unable to determine from his report whether the inductive-deductive method was statistically superior to the deductive-descriptive method or whether the difference was due to chance factors; no statistical data were reported.

Boeck's design has little, if any, external validity outside the experimental school. The inductive-deductive method was not employed outside the experimental school. One should only consider his findings as tentative.

## Transfer

Seven comparative research studies have been identified that investigated the transfer of science training. These studies are listed in Table 8. Only the studies by Judd (1908) and Hendrickson and Schroeder (1941) reported significant findings concerning transfer, both favoring the expository methods.

Table 8

Significant Findings Reported by Experimenters for Transfer of Science Learning

| Study | Year | Level | Method Favored | Subject Area |
|---|---|---|---|---|
| Judd | 1908 | Elem | Expository | Phy Sci |
| Hendrickson and Schroeder | 1941 | Jr H | Expository | Phy Sci |
| Craik | 1966 | Coll | Neither | Zoology |
| Dennison | 1969 | Jr H | Neither | Gen Sci |
| Tanner | 1969 | Jr H | Neither | Phy Sci |
| Babikian | 1970 | Jr H | Neither | Gen Sci |
| Chambers | 1971 | Coll | Neither | Phy Sci |

Transfer was measured in various ways. For the study by Hendrickson and Schroeder (1941), transfer was defined in terms of improvement from the first task to the second task. Chambers (1971) used trials to criterion or errors-to-criterion on the transfer task as the dependent variable. Other studies, e.g., Tanner (1969) and Babikian (1970), used raw scores, obtained from written tests, to measure transfer.

Both immediate and delayed transfer measures were used by Dennison (1969), Tanner (1969), and Babikian (1970). In each case, the delayed measure was administered four weeks after the administration of the initial posttest.

The study by Craik (1966) used the A.C.E. *Test of Science Reasoning and Understanding*. It is debatable whether this test can be considered a transfer test; Craik did not indicate that it was.

Judd (1908), using two groups of elementary school students, studied the teaching of a relationship between the depth of water and the refraction it produced. One group was taught, in verbal form, the principle of refraction. They then practiced throwing darts at a submerged target. The other group received no verbal instruction concerning infraction; they used their instructional time practicing throwing darts. Judd found that the group receiving verbal training performed better on a transfer test with a change in depth of the water. Judd's report contains no statistical data. It is therefore impossible to determine if his findings are in keeping with the data. Furthermore, Judd's experiment contains confounding psychomotor variables and novelty effects. This study was included for review primarily for historical purposes.

Hendrickson and Schroeder (1941) conducted a study patterned after Judd's (1908) experiment. Ninety boys in eighth grade served as the subjects. Thirty were assigned to each of three groups: control, experimental group A, and experimental group B. Group A was told an elementary explanation of refraction. Group B was given the same explanation plus a statement of the fact that the depth of water changed the amount of refraction. The control group (pure discovery) was given no instruction concerning refraction. An air gun was used with a horizontal target in a 20-gallon tub with the bottom diameter of 20 inches, a top diameter of 24 inches, and a depth of 11 inches. The subjects fired from a platform 18 inches high and at a distance of eight feet from the center of the target. All of the work was conducted after school hours with each student working privately. A student was considered successful when he scored three consecutive hits. A record was kept of the number of hits to criterion. Each student was required to work on two problems; the first problem used a water depth of six inches and the second problem used a water depth of two inches. The students were required to satisfy the criterion of three consecutive hits for the first problem before working on the second problem. Transfer was defined in terms of the improvement from the first task to the second task.

Based on an examination of group means and standard deviations, the experimenters concluded that, for both tasks, Groups A and B learned more rapidly than the control group, and group B learned more rapidly than group A. Transfer occurred for all groups with group B improving more than group A, and group A improving more than the control group.

The study by Hendrickson and Schroeder did not rely on a systematic procedure for testing their hypotheses, i.e., no decision rules were employed for testing their hypotheses. Furthermore, their study contains confounding psychomotor variables.

Tanner (1969) compared three methods (expository, guided-discovery, and unsequenced-discovery) of programed instruction for teaching the principles of simple machines to 389 ninth-grade students in fourteen general-science classes at a single high school. Tanner, in describing the instructional programs, stated:

> The self-instructional programs of this study used principles derived from the basic work principle of simple machines as it is presented

> in most eighth- or ninth-grade science texts. The frames exemplified the principles, and consisted of diagrams of various levers, inclined planes, and wheel-and-axle systems. The learner responded to each frame by computing certain weights or distances missing from the diagram. After recording his answer he received feedback as to the correct response. In the expository-deductive program, each group of frames was preceded by an illustrated explanation of the appropriate principle. The discovery-inductive and unsequenced-discovery programs contained only the frames: the learner had to infer the principle if he were to attain them at all. Guidance was further reduced in the later program through a random ordering of the frames. (p. 137)

Comprehension, lateral transfer, vertical transfer, and retention measures were used to evaluate the relative effectiveness of the treatments. Concerning the transfer measures, Tanner stated:

> . . . The transfer measures required behaviors not required by the program. The vertical transfer measure presented the same machines, but the learner had to combine the programed principles in various ways in order to respond correctly to the items, i.e., he must form higher order principles. The lateral transfer measure required transfer to new machines. (p. 138)

Odd-even reliability estimates for the various instruments ranged from .76 to .93.

The three forms of the program were randomly distributed, in approximately equal numbers, within each of the 14 classes. An uninstructed control group was included in the experiment. Ninety-seven subjects were randomly selected from the three treatments to take the retention measure four weeks after instruction without having taken the comprehension and transfer measures immediately after instruction, providing retention performance that was not influenced by previous practice on the criterion measures.

The data were analyzed using ANOVA and ANCOVA; the covariate was the scores on the mechanical reasoning subtest of the DAT, which was administered two weeks prior to the start of the experiment. Tanner reported no main-effect differences on any criterion measure. It was found that the experimental groups, combined, were not superior to the control group on the transfer measures. It was found that girls and high-intelligence students tended to perform best after using discovery treatments, whereas boys and low-intelligence students tended to perform best after using the expository program.

The discovery treatments may have suffered because of the lack of familiarity with the technique by students involved with this method and mode of presentation.

This study appears to have been well designed.

Babikian (1971) conducted a six class-period study to determine the effectiveness of teaching six principles involving buoyancy in liquids to eighth-grade students by the discovery, laboratory, and expository methods. The expository method was characterized as

> . . . a method of instruction in which the teacher presents the science concepts to the students verbally.
> (a) The concept is stated first, and then examples are given for further clarification.
> (b) The teacher may make use of no A-V material but the charkboard.
> (c) Students are allowed to ask questions or to discuss the concept. (p. 201)

In the laboratory method

> . . . the teacher presents the concepts, as well as the procedural instructions for their verification in a printed laboratory manual, and provides each student all the equipment necessary for verifying each of the individual concepts.
> (a) The concept is stated first, and then the procedure is described for the verification of the concept.
> (b) The teacher may explain the concept if necessary.
> (c) Students may ask questions.
> (d) Students do not cooperate with each other.
> (p. 201)

The discovery method was characterized as

> . . . A method of instruction in which the teacher presents in a printed manual the procedural instructions for the discovery of an unstated concept, and provides all the equipment necessary for each subject to discover the concept himself.
> (a) The teacher may explain the procedure but not the concept.
> (b) Students may request assistance on procedural matters.
> (c) Students may inquire about the concepts being discovered, but they get only "yes" or "no" answers from the teacher.
> (d) Students do not cooperate with each other. (p. 202)

Over a three-week period, the investigator taught nine classes from a single junior high school; a separate method was thus used to teach three classes each week. No steps were taken to insure that the investigator was faithful to the prescribed teaching methods. Further, by teaching the three discovery classes the first week, the three laboratory classes the second week, and the three expository classes the third week, experimental bias may have been built into the study. Also, since all students were from the same school, the study may have been contaminated by communication among students. Two hundred and sixteen students were randomly assigned to the three treatments, and each group was further divided both according to two I.Q. levels and by sex, resulting in nine classes of 24 students each.

A posttest was given to evaluate the effectiveness of the three treatments, immediately after instruction and four weeks later. A pretest was given to assess the students' previous knowledge about the learning tasks. There is the possibility that the pretest may have sensitized the students to the content of lessons, thus confounding the study. The posttest consisted of 38 items and measured five areas, including retention and transfer. A reliability coefficient of .76 was reported for the test; approximately 25 per cent of variance in test scores was due to error variance. A control group was not included in the design of the experiment; thus any differences found may be due to maturation factors alone.

The data were analyzed by employing a 3x2x2, treatments-by-level-by-sex, ANOVA design with 18 scores per cell. The investigator gives no rationale for pooling the data into this kind of a design; the classes were taught in nine classes of 24 students each. The experimental unit for the study was the student score, a questionable move. After significant F-values for treatment effects, pairwise t-tests were employed at the .01 level of significance for explaining differences between means. It is well known that such a procedure exploits many of the chance differences and leads to an inaccurate probability model against which decisions are based.

Concerning transfer, the treatment effect was significant at the .05 level of significance, but not at the .01 level which was established as the experimental standard. No power levels were reported for any of the significance tests.

Due to a weak experimental design and a questionable data analysis, the findings from this study do not support a superior teaching method for retention or transfer of science training.

Chambers (1971) studied in the effects of discovery learning and over-learning on transfer power. Fifty-six college students were involved in the study, and the serial anticipation task involved learning a principle relating the name of a lens with the amount and direction of distortion it produced in the apparent location of an object. Four different lenses were used in the experiment. Prior to the experiment, training sessions were held to familiarize the students with the techniques and apparatus.

A 4x2 factorial design was used with four levels of discovery (didactic, discovery, and two guided-discovery treatments) and two levels of over-learning. Chambers does not explain the differences between the experimental treatments. Over-learning consisted of an additional practice trial for each two trials required to attain the criterion. The criterion were five successive, correct anticipations.

One week after instruction, a transfer task was administered, and the number of errors to criterion of five successive anticipations was used as the dependent variable. The transfer task involved an extention of the principle learned in the experiment; it was not stated what this extention consisted of, and no reliability data were reported. Chambers reports no statistical data in his report. He concluded that the overlearning group was significantly (.05 level) superior on the transfer measure to the other group. There were no significant differences among the four treatments, and interaction was not present. Chambers interprets his findings to mean that practice is more important than discovery in inducing transfer.

From Chamber's report of his experiment, it would be difficult, if not impossible, to replicate his experiment. Further, no statistical data were reported; it is, therefore, impossible to determine whether his findings are in keeping with the data.

No superior method emerges after reviewing and analyzing the comparative research studies on transfer of training in the area of science.

## Industrial or Vocational Education

Seven studies have been identified that investigated the effects of retention or transfer of industrial or vocational training. These studies are listed in Table 9. Since only three of these studies reported significant findings concerning retention or transfer, they will be analyzed jointly for retention and transfer. Although they reported no significant findings concerning retention or transfer, the studies by Grote (1960), Moss (1960), and Tomlinson (1962) will be reviewed for illustrative purposes.

A series of five related studies (Ray, 1957; Rowlett, 1960, 1964; Moss, 1960; and Grote, 1960) have compared the relative effectiveness of direct-detailed and directed-discovery methods of teaching technical subject matter from the area of industrial education. Presumably, technical subject matter was chosen because of its unfamiliarity to the learner.

In each study, approximately one hour of instruction was presented by means of tape recordings integrated with student workbooks. The studies by Ray, Rowlett, and Moss employed instructional aids or models. All studies used subjects stratified into three ability levels to test for levels-by-treatments interaction. All studies included an uninstructed control group which took all criterion tests.

Essentially, in the direct-detailed method, the learning task was presented in a detailed, step-by-step procedure, where leading questions and sequences of questions, each followed by a pause, were used to direct the student's attention to problems or applications to be discovered by the students.

## Table 9

### Significant Findings Reported by Experimenters for Retention or Transfer of Industrial or Vocational Training

| Study | Year | Transfer | Retention | Level | Area |
|---|---|---|---|---|---|
| Ray | 1957 | D* | D | Jr H | Micrometer Skills |
| Grote | 1960 | N** | N | Jr H | Mechanics (Physics) |
| Moss | 1960 | N | N | H S | Printing |
| Rowlett | 1960 | D | D | Jr H | Orthographic Projections |
| Tomlinson | 1962 | N | N | H S | Metallurgy |
| Rowlett | 1964 | D | D | H S | Orthographic Projections |
| Luck | 1966 | N |  | H S | Automotive Topics |

*Discovery method.

**Neither method favored.

For each study, five criterion measures were employed: initial learning, early and late transfer, and early and late retention. The studies differed in methods of experimental analyses, and each study used the student score as the experimental unit, rather than the class mean. The tape recorded instruction was group presented, and there is the possibility that a student's non-verbal behavior, when manipulating models, could have provided clues to other students. Indiscriminate pooling of the data was also common to these studies; the experimental conditions for learning within each class were not completely controlled, e.g., time of day, physical setting, and random external events, rendering different treatment effects under the same method.

Ray (1957) studied the effects of teaching micrometer skills and principles to 117 ninth-grade boys from three junior high schools using two methods (direct-detailed and directed-discovery) of tape recorded instruction. The tape recorded instruction was supplemented by presenting illustrations through the use of 35 mm. slides. Students in the directed discovery method, by studying illustrations and manipulating micrometers, were to discover the principles and skills regarding the micrometer without direct instruction or assistance from the experimenter or other students. Leading questions were sometimes presented by the experimenter.

Students were taught in groups of nine, each group consisted of three students from each of three ability levels determined by I.Q. scores. The effectiveness of instruction was evaluated by an initial learning test, a one-week retention test, a one-week transfer test, a six-week retention test, and a six-week transfer test. The tests had reliability coefficients ranging from .92 to .97, and all tests were of the multiple choice type. Ray concluded that the directed-discovery treatment was significantly superior to the direct-detailed treatment on both transfer measures and on the six-week retention measure. No other significant findings were reported.

Prior to instruction, Ray conducted six randomized, one-way ANOVA tests to be certain that the three treatments were matched according to I.Q., that the treatments-within-levels were matched with respect to I.Q., and that achievement scores and ages were matched across treatments. In every case, a F-ratio less than one was calculated; in fact, every ratio was significantly small and suggests that the model underlying the analysis of variance has been violated in some way according to the criteria established by Meyers (1966, pp. 66-67).

For each criterion measure, the data were analyzed using a 2x3, treatments by levels, ANOVA model. The control group was not used in testing the experimental hypotheses. The unit of analysis was the student score, a questionable procedure. Each cell in the 2x3 design had a total of 15 students, but the students were taught in groups of nine at each of three schools. Ray gives no justification for this pooling of scores for each criterion measure.

Certain evidence of confounding can be found in Ray's study. At the outset of the experiment, the experimental students were told that they were participating in an experiment, thus invoking the Hawthorne effect. No procedures were taken to control the novelty effect of the instruction. Further, experimental bias could have been generated since students were not allowed to ask questions; the students in the discovery classes, because of the presentation, might have been at a slight advantage for resolving conflicts. Ray assumed that the method of presentation, including instruction, would be understood by all students. There is also the possibility that some students were not given equal opportunities for learning because of audio or visual handicaps. Ray's findings should only be considered as tentative.

Rowlett (1960) extended Ray's (1957) study by comparing the direct-detailed and directed discovery methods of teaching orthographic skills and principles to 168 ninth-grade students from a single high school. For both methods, instruction was provided by a tape recorder which was keyed to a workbook. Students were also given three small models (blocks) to manipulate. No feedback of information to the students was provided for.

Three ability levels were established by using a standardized intelligence test. Forty-nine minutes of instruction was presented to six groups of 24 students, equally and randomly assigned in terms of sex and ability level. Four students were in each sex-by-ability level cell. A control group consisting of 12 boys and 12 girls representing the three ability levels received no instruction but took the criterion tests.

Provisions were made to control the mortality effect by exposing 12 additional students to the treatments under experimental conditions. An initial-learning test was administered immediately following the instructional period, and retention and transfer tests were given to all students at twelve days and six weeks after instruction. Both retention tests were identical to the initial-learning test, and both transfer tests were identical. All tests were of the multiple choice type. Split-half reliability coefficients were reported and ranged from .86 to .94.

For the I.Q. measure, the data were pooled into a 3x2, levels-by-treatments design. No justification was given for this regrouping of the data. Each pooled group had an equal number of male and female subjects; thus, each ability level consisted of two subgroups containing the same number of elements.

In order to show that each subgroup within ability level was matched according to I.Q. scores within treatments, 10 one-way, ANOVA significance tests at the .05 level were conducted; each significance test yielded no significant results, according to Rowlett. Assuming independence of comparisons, the probability of one or more comparisons being uncorrectly rejected at the .05 level is .40 $(1-.95^{10})$. One of these tests, an ANOVA of I.Q. scores of students by treatments, yielded an F-ratio of .008 which is significantly small under the criterion established by Meyers (1966, p. 66-67) and suggests that the model has been violated. Thus, the design was considered as consisting of six matched pairs of subgroups, two at each level. For each criterion measure, the mean score was computed for each pair of matched subgroups; the data were then analyzed by using a t-test for correlated samples with five degrees of freedom.

To test for linear interaction between treatments and ability levels for each criterion measure, Rowlett used 15 t-tests (three for each measure) and 15 F-tests (three for each measure); the t-tests were used to establish homogeneity of means, and the F-tests were used to establish the homogeneity of variance assumptions underlying the t-test. Two F-tests were found significant at the .10 level, and one t-test was found significant at the .05 level. Assuming independence of tests, by chance alone, these would be expected to occur.

Rowlett concluded that the evidence indicated that no interaction between methods and ability levels was present for any of the five criterion measures. He also concluded, among other things, that the directed-discovery treatment was superior to the direct-detailed treatment on both transfer tests and on both retention tests.

Several forms of contamination appear in Rowlett's study. The students were told that they were participating in an experiment. No attempt was made to control the novelty effects. Since the students were all from one school, it is conceivable that students communicated during the experiment. Further, since no feedback was provided for the learner in the lesson, one cannot be sure whether the student discovered concepts or skills as a result of the lesson or as a result of some hint or phrase given at a later date. Confusion on the part of the students may have inhibited future performance.

Because Rowlett was concerned with methods-by-ability interaction, a two-factor ANOVA model would have been more appropriate to analyze his data and would have controlled the experimentwise error rate. It appears that MANOVA models were generally not well known in the early 1960's.

As a result of Rowlett's questionable data analysis, his findings should only be considered as tentative.

In a similar study, Rowlett (1964) measured the effectiveness of the same two methods of instruction as measured by five criterion tests in a 43-minute learning situation involving the same learning materials, but with nine models (blocks) instead of three. One hundred and forty-seven female students were involved in the study and were randomly selected from non-freshmen students enrolled in a required social science course at a state college. The students were randomly assigned to three levels of ability within methods. A control group was considered in the design and received no instruction.

All instructed students were again told that they were taking part in an experiment, thus invoking the Hawthorne effect. Students were taught in six groups of 21 students. Each group was composed of seven students from each the high, average, and low ability levels as determined from scores on The Revised Minnesota Paper Form Board Test.

The data were analyzed by using a treatments-by-levels design with 21 students per cell, a different arrangement from the manner in which they were taught. No explanation or justification was given for this rearrangement of the data. For each of the five criterion measures, an F-test (for homogeneity of variances) and a t-test for uncorrelated groups was conducted, resulting in a total of 10 significance tests and a questionable experimentwise error rate.

The experimental unit used in the analysis of the data was the student score; this is questionable since it is not clear that the student scores were independent. To test for interaction, six significance tests were conducted for each criterion measure, a total of 30 significance tests for studying interaction. Ten pairwise t-tests at the .05 level of significance were conducted and showed that both experimental groups were superior to the control group on each criterion measure. This is not an appropriate procedure for comparing a control group with two experimental groups. At least 50 significance tests were conducted during the experiment, raising the experimentwise error rate well beyond the .05 level.

Rowlett concluded that (1) the directed-discovery group was superior to the direct-detailed group on the six-week transfer test, (2) there was no evidence of interaction, (3) the average ability directed-discovery students were superior to the average ability direct-detailed students on the six-weeks retention test, and (4) the low ability directed-discovery students were superior to the low ability direct-detailed students on the six-weeks transfer measure.

Due to a questionable analysis of the data and evidence of contamination of the treatments, Rowlett's findings should be considered as tentative. A factoral ANOVA design would have been more appropriate for checking the presence of interaction.

Grote (1960), in an effort to extend the results of Ray (1957) and Rowlett (1960), compared the direct-detailed and directed-discovery methods of taped instruction. He taught selected principles of mechanics and their applications to groups of simple machines to 180 eighth-grade students from a junior high school. Grote failed to control the Hawthorne effect by informing the students that they were participating in an experiment.

A sex-by-levels-by-treatment, factorial design was employed in the experiment. Three ability levels were established on the basis of SCAT total raw-scores. Based on the SCAT scores, Grote concluded that he was working with an atypical population. There were two instructional sessions, using different tasks, spaced eight days apart, and providing approximately 39 minutes of instruction during each session. Each session involved non-manipulative technical material. Lesson one dealt with the lever while lesson two dealt with the pulley and wheel and axle machines. A control group participated in the testing program but received no instruction.

After the first instructional period, the two experimental groups were each subdivided into two groups, each subgroup of students was assigned to one of the two treatments for the second lesson. Six initial-instruction groups were each composed of 24 subjects who had been randomly assigned on the basis of sex and ability levels. For the second lesson, the six instructional groups were randomly regrouped to form eight groups of 18 students each. Treatments were not randomized as to the day of week for either learning session. Failing to control this variable may have had some concomitant effect on the experiment. Further, this regrouping and changing of environment may also have led to confounding.

A total of six multiple choice, power, criterion tests were administered during four testing sessions. Initial-learning tests were given after each instructional period, and a combination retention-transfer test was administered at one and six weeks after the second instructional period. Subjects were notified in advance about subsequent tests, thus encouraging communication between students. Split-half reliability coefficients for the two initial-learning tests were .76 and .67.

The data were analyzed using a 2x3x2, sex-by-levels-by-treatment, ANOVA design with six subjects per cell. In so doing, Grote possibly confounded the analysis in two ways. First, the student score was used as the unit of analysis, whereas the instruction was group presented with possible communication between students. The scores, therefore, may not be independently distributed. Secondly, the scores were pooled or regrouped for each criterion measure, without justification, from the original design in which they were taught in order to obtain a cell size of six in the 2x3x2 design.

Inappropriate multiple comparisons were used on two separate occasions. To show that the treatment groups were superior to the control group on initial learning measures, pairwise t-comparisons were employed instead of ANOVA. Further, because Grote did not employ appropriate post hoc comparisons following a significant F, he inefficiently employed 35 2x3x2 ANOVA designs to evaluate learning on each of the criterion measures. At least 35 significance tests were made, and if they were independent, which they were not, two type I errors would have been expected to occur by chance.

Grote concluded, among other things, that (1) the direct-detailed technique was significantly superior to the directed-discovery technique on the initial-learning test after the first session, (2) on the initial-learning test after the second session, there were no significant differences between the experimental groups, and (3) the sequence of directed discovery followed by direct-detailed instruction was the most effective as measured one week after instruction for retention and transfer.

Grote's experiment contains many sources of possible contamination. For the second learning session, subject-matter content and treatment variables confound each other, possibly canceling effects. Since all students were from one school, it is very likely that communication existed between students during the study. Grote's findings should be considered as tentative.

Moss (1960), extending the work of Ray (1957) and Rowlett (1960), taught the content of letterpress imposition to 106 (originally 108) high school students, all from the same high school, using the direct-detailed and directed-discovery methods. A 3x3, treatments-by-levels design was employed in the experiment. A control group was used.

Three ability levels were formed on the basis of I.Q. scores. The mean Otis I.Q. score for the experimental population was 95.3, and there was some evidence to suggest that these scores were not normally distributed. To determine whether the I.Q. scores were distributed equally among the treatment groups, a simple ANOVA randomized design was employed. Based on 2 and 103 d.f., an F-value of .12 was obtained, which is significantly low and implies that the

ANOVA model has been violated in some fashion. This is not surprising since the three groups were of unequal sizes and constituted a non-normal distribution. Although the data were analyzed by using a 3x3 factorial design with approximately 15 students in each treatment-by-level cell (6 cells) and three control cells (one for each level), the treatments were presented to four groups of students, two groups received each treatment. A fifth group consisted of all subjects assigned to the control group. This pooling of data is a questionable move, particularly since the lesson was presented in group format.

The effectiveness of the treatments was determined from scores made on initial learning, retention, and transfer tests administered after the formal instructional period. Stability and equivalence reliability coefficients ranged from .57 to .77 for the two instruments used in the evaluation.

Five 2x3 (treatments-by-levels), fixed effects ANOVA models were used to test the effectiveness of the treatments for each of the criterion measures. The student score was used as the experimental unit, a questionable procedure. For each criterion measure, at least one significantly small F-value was calculated, indicating that the ANOVA model had been violated in some manner. Moss reported no significant findings.

Tomlinson (1962) compared the relative effectiveness of inductive, deductive, inductive discovery, and inductive-discovery-confirmation methods of presenting certain content on the metallurgy of carbon steel to 162 junior and senior high school students from a single school. The lessons were presented in written, programed form and contained 10 sketches. The inductive method (A) was defined as:

> . . . a relatively straightforward presentation of written material, divided into eight sections. Each section of the material closes with an underlined summary statement of the broad generalization (s) developed within the sections. The total learning passage is cumulative, closely interrelated and interdependent. (p. 4)

The inductive-discovery-confirmation method (B) presented the eight sections and accompanying sketches as in the inductive method. Further,

> . . . a general question is presented at the end of each section to stimulate the learner to synthesize (discover) a generalization(s) appropriate to the content of the section. If, in fact, he does this, then he may confirm his generalization statement of a relationship by comparing it to that on a check sheet where the summary statement or broad generalization(s) developed by the experimenter is given. The subject may consider various alternatives in forming his generalization or in the process of comparing his to the one formulation presented. (p. 4)

In the deductive method (C),

> . . . each section of the learning passage is introduced with the broad generalization(s), the same state that is used for the summary in Treatment A. The content of each section, identical to the other treatments, is conceived to be a detailed and specific development within the more abstract and inclusive generatlization(s) presented as in the introduction. (pp. 4-5)

For the inductive-discovery method (D)

> . . . The eight sections, identical to the other treatments, close with the same broad question (as in treatment B) to stimulate the subject in forming his generalization(s) or synthesis of the content presented. The learner is left to his own resources and may proceed at his own pace and level. (p. 5)

A control group received no formal instruction, but received all criterion measures.

Three ability levels were established, and students were randomly assigned by classes and levels within each treatment, resulting in 36 students exposed to each treatment. As a pre-organizer, monitors read a common orientation to each experimental group at the beginning of the initial session. The students were allowed 45 minutes to study, from individual booklets, a written passage organized into eight interdependent sections. Following the individual study, a 15-minute structured group-demonstration was presented by the monitor in charge of the treatment session as a post-organizer to illustrate the principles and relationships presented in the written lesson.

The effectiveness of instruction was evaluated by scores made on measures of initial learning, one and five weeks retention, and one and five weeks transfer tests. The measures employed subtests containing true-false, multiple-choice, and free-response type questions. The stability coefficient of reliability for the true-false and multiple-choice measures ranged from .41 to .49 and .64 to .74, respectively. The stability reliability coefficient for the free-response measure was found to be .67.

A 2x3x2, class-by-level-by-treatment, ANOVA design was applied separately and independently to scores for each combination of treatment pairs, resulting in six factorial designs for each of the ten criterion measures, or a total of 60 separate analyses. Assuming independence of tests of significance and an alpha level of .05, three analyses should result in a type I error by chance factors alone. For each analysis, the student score was taken as the experimental unit, a questionable move. To detect significant differences in mean scores between the control group and each of the experimental groups for the two tests of initial learning given immediately after instruction, t-tests were conducted.

This leads to an inaccurate probability model against which decisions are based. Further, multiple comparisons tend to inflate the experimentwise error rate.

Tomlinson concluded, among other things, that (1) all methods were equally effective when success is measured in terms of one-week retention and transfer tests, (2) the inductive-discovery-confirmation method is inferior to all other methods when success is measured in terms of retention and transfer at five weeks after instruction, and (3) an expository method, inductive or deductive, stating the generalizations is superior to the methods including questions to stimulate the student to form his own generalizations when success is measured by retention and transfer at five weeks.

Tomlinson's use of the post-organizer and his lack of control over the Hawthorne and novelty effects tend to confound the study. If differences did exist, it would be difficult to argue that they were caused solely as a result of the treatment. Further, this experimental design has limited external validity. A poorly controlled design, tests of questionable reliability, and a questionable analysis of the data make it risky to draw any substantial conclusions from this experiment.

In summary, no conclusive evidence emerges from the above industrial or vocational studies concerning a superior method of teaching with respect to retention or transfer measures.

## Geography and Language

Seven comparative research studies have been identified, five in language arts and two in geography, which have compared the effectiveness of discovery and expository teaching strategies by retention or transfer of training measures. These studies are listed in Table 10. Only the study by Rizzuto (1970) reports significant findings concerning retention or transfer. The studies by Wiesner (1971) and Lahnston (1972) will be reviewed and analyzed for illustrative purposes.

Rizzuto (1970), in a seven-week study, compared the relative effectiveness of inductive-discovery and deductive-expository methods of teaching concepts of language structure to 165 eighth-grade students from a single school. A five-week instructional period contained 20 45-minute sessions. The inductive method presented a loosely structured sequence of specific examples from which the learner was to discover and verbalize the concept. Rizzuto stated:

## Table 10

## Significant Findings Reported by Experimenters for Retention or Transfer of Geography or Language Learning

| Study | Year | Transfer | Retention | Level | Area |
|---|---|---|---|---|---|
| Naughton | 1962 | N** | | Elem | Spelling |
| LaRocque | 1965 | N | N | Jr H | Figurative Language |
| Britton | 1969 | N | N | Elem | Phonics |
| Rizzuto | 1970 | D* | D | Jr H | Language Structure |
| ≠Hermann | 1971 | N | N | Jr H<br>Elem | Geography (map reading) |
| Wiesner | 1971 | N | N | Jr H | Spelling |
| Lahnston | 1972 | N | N | Elem | Geography (transportation principle) |

*Discovery method.

**Neither method favored.

≠Students were given rules.

> At the outset, the teacher had only to pose the linguistic topic of the lesson and focus the learner's attention on the problem. As the lesson progressed, the teacher asked open-ended unanswered questions which prompted and guided pupils in making discoveries. When necessary, the teacher supplied information, but communicated among students was emphasized rather than teacher-student dialogue. (p. 270)

The deductive method presented a didactic, verbal exposition of the concept which the learner was to verbalize and apply to examples. Students were stratified by sex and ability level and were randomly assigned to the two treatments; each treatment had three classes assigned to it. A control group was included in the design and received only the criterion measures. Six teachers were randomly selected and assigned, one to each class.

To verify teacher adherence to the treatment, each experimental teacher was video taped on two occasions. This author finds it difficult to understand how this procedure would insure teacher fidelity to treatment; surely the teacher knew that he was being taped.

A test measuring recognition and transfer was administered on two separate occasions, immediately after instruction and two weeks later. The strategy employed in the transfer measure was the use of nonsense sentences for which the students had to transfer concepts of language structure. Both subtests were multiple-choice and had reliability coefficients of .88 and .86.

The data were analyzed using six 2x3x2 ANOVA designs, each at the .05 level of significance. The student score was used as the experimental unit, a questionable procedure. Dunnett's test was used for comparing the control group mean with the experimental group means; in all cases, the experimental group means were significantly higher than the control group mean.

Rizzuto concluded, among other things, that the inductive method was significantly (.01 level) superior to the deductive method on all six criterion measures (recognition, transfer, and total score for each testing period). To some extent, this is not surprising since the intercorrelations between the recognition and transfer sections of the immediate and delayed tests were .78 and .81, respectively. Approximately 64 per cent of the variance in scores is common variance which indicates that both tests are sampling, to a sizeable degree, the same behavior.

Rizzuto's experimental design has limited external validity. Pooling of the data and the choice of experimental unit are both open to controversy. Further, from his statistical summary, it is impossible to determine if his findings are in keeping with the data. Therefore, the results of his study should only be considered as suggestive.

Wiesner (1971) conducted a six-week experiment to test the effectiveness of discovery versus expository methods and teacher-guided versus independent procedures for teaching six spelling principles to 348 sixth-grade students. The teacher-guided procedure was to allow for maximum group discussion, and it is not clear, from Wiesner's report, just what the independent procedures were. Further, no definitions were reported for either teaching method used in the study. An experimental program was written for the study and consisted of six spelling lessons. Sixteen intact, sixth-grade classes in a public school system were used in the study; four classes were randomly assigned to each treatment. It is questionable that the intact classes were representative random samples since few classes are formed without some intentional selection process. A control group was not included in the experimental design. No mention was made of controlling the teacher variable. Care was taken to control the Hawthorne effect. Wiesner (p. 218) stated, ". . . teachers were cautioned against the use of any procedures which might suggest the experimental nature of the program or bias the results."

One immediate and one delayed (six-weeks) posttest were given to determine achievement in terms of retention, transfer, and problem solving ability; each test consisted of three subtests. The retention subtest consisted of words which had been practiced in the lessons, whereas the transfer subtest consisted of words which had not been practiced, but to which the principles applied. Concerning test reliabilities, Wiesner states,

> . . . Reliability was checked by giving the test to four classes which were not otherwise involved in the study. The resulting coefficients of equivalence were significant beyond the .01 level. (p. 218)

This author does not understand how this procedure establishes the reliability of the test for its intended use.

In a questionable procedure, the data were analyzed by employing a 2x2 analysis of variance model. No summary table or other statistical data were included in the report. Wiesner reported no significant findings. Inadequate reporting makes it difficult to be certain just what happened and how the data were analyzed. It is, therefore, impossible to determine whether her conclusions are in keeping with the data. This is a poorly reported study, and, as a result, it would be impossible to replicate.

Lahnston (1972) investigated the effects of a demonstration-deductive-expository strategy and an inductive-discovery strategy for teaching a geographic generalization and its component concepts to third-grade students. Twenty-four subjects were classified in two ability groups and then randomly assigned to one of two treatment groups. The two ability levels (the top and bottom 30 per cent of the I.Q. scores) were randomly chosen from the total third grade membership at a certain school. The generalization used for the learning task was "The sites of the cities are often places where goods are transferred from one means of transportation to another," and the component concepts consisted of the means of transportation, transfer of goods, and the site of a city. In both treatments, a slide projector was used to present concrete examples. Instruction was individualized and consisted of five 15-minute lessons, each administered on consecutive school days. Students were randomly assigned to one of three teachers.

The expository strategy began with the teacher stating the generalization and explaining it by offering an example. This was followed by visually illustrating the generalization, defining and illustrating the constituent concepts, identifying the attributes of the concepts, and presenting positive and negative examples of the concepts.

The discovery strategy began by presenting pictorial examples to the student and having him identify the features he observes. The teacher then made a list of the student's observations, after which the student was given his list of terms and asked to group together similar items and label categories. He was then encouraged to construct and test a definition for each category. After doing this, the student was asked to compare and contrast a series of maps and other pictures; these similarities were then entered onto a data retrieval chart. From this data, the student was asked to generalize his findings.

Five criterion measures were used in the experiment: retention (initial achievement), immediate transfer, delayed retention, delayed transfer, and trials to mastery. On the sixth school day after the start of instruction, students were given an immediate posttest consisting of 20 items, 10 achievement and 10 transfer. One day after the posttest, each student in both treatments received 10 minutes of additional instruction, consistent with their original treatment format and designed to help the student reach mastery. If a student achieved a 90 per cent correct score on the mastery test, the instruction for him stopped. A maximum of three additional trials to reach mastery was available to each student. Immediately after each trial, a mastery test of 10 items (5 achievement and 5 transfer) was given. If a student failed to achieve mastery on the third mastery trial, one additional mastery trial was assigned to him. Points were assigned for mastery trials; one point for mastery on the first trial, two points for mastery on the second trial, three points for mastery on the third trial, and four points for failing to achieve mastery. A delayed posttest with two subscores, retention and transfer, was administered two weeks after the student's last mastery lesson.

Five tests were constructed to evaluate the relative effectiveness of the two strategies: one immediate posttest, three mastery tests, and one delayed posttest. The reliability coefficient for each posttest subtest was at most 0.43. This resulted, in part, from the subtests containing only 10 items.

The data were analyzed by using five 2x2, ANOVA designs. For each design, the student score was used as the experimental unit. Each cell contained a total of six scores.

To show that there were no differences in ages or I.Q.'s for each factor, t-tests were used.

Lahnston concluded that the students in the expository strategy scored significantly (.01 level) higher than the students in the discovery strategy on the initial-achievement measure. No other significant differences were reported.

The F-values computed for both transfer subtests were significantly low as judged by Myers criteria.

> . . . The occurrence of F's so small that their reciprocals are significant or the occurrence of many F's less than one in a single analysis of variance merits further consideration. Such findings suggest that the model underlying the analysis of variance has in some way been violated. (pp. 66-67)

Evidently, the treatments contained some systematic factor that made the groups more homogeneous.

By exposing the students to mastery trials following the first posttest, Lahnston confounded the study. One cannot be sure that the mastery trials are not contaminating the treatment effects. These two variables need to be studied separately under similar conditions to determine their relative effectiveness on achievement, retention, and transfer measures.

## General Summary

This review of the comparative, experimental research done on discovery teaching and discovery learning in the areas of science, mathematics, industrial education, language, and geography has failed to identify a superior teaching method with respect to transfer or retention measures.

The reader of this study may have formulated a sense of dissatisfaction with the state of experimental research in the area of discovery teaching and in education in general. He may also come away with great cynicism and feel that none of the studies reviewed are worth any consideration. But, this study should not be regarded as a condemnation of professional and non-professional research, especially in the area of "discovery" research. Rather, it should be taken as an effort to bring to light the research activities in discovery teaching and to improve the quality of experimental research in education. It is very easy to criticize and find weaknesses in another's work, but it is difficult to produce flawless work on one's own. An experimenter is often too involved in his research to recognize defects or inadequacies in his design or data analyses, while the impartial observer, not being personally involved, may readily recognize errors and weaknesses that occur. Also, it should be kept in mind that sophisticated, statistical technology was generally not available to the average researcher in education prior to the 1960's.

In educational-methodology studies, the risks of arriving at incorrect conclusions are seldom, if ever, damaging. If a study erroniously suggests that one method is better than another for teaching certain subject matter, the learner is not at a disadvantage because both methods (such as discovery and expository) are usually effective for learning. Thus, if a teacher erroniously accepts one of the many significant findings reported and modifies his teaching behavior as a result, probably little, if any, harm should result to his students.

This study has identified a large number of teaching strategies, both discovery and expository, that are effective for teaching a variety of subject-matter content. The strategies and ideas involved within the many studies reviewed should provide suggestions and guidance for both the inservice and preservice teacher. This author feels that this identification is one of the major contributions of this study.

This study should also do much to strengthen experimental research in education for the future. Many pitfalls have been identified, and many research hypotheses have been pointed out, still needing to be decided. Chapter V is devoted to suggestions for strengthening experimental, methodology research in education, while Chapter VI discusses problems and hypotheses arising from this study for future research.

# CHAPTER V

## RECOMMENDATIONS FOR IMPROVING FUTURE RESEARCH

As was pointed out in Chapter III and exemplified in Chapter IV, many problems exist in conducting experimental, methodological-research studies. Some of the more salient problems are:

1. not specifying *a priori* alpha levels.
2. not reporting power levels for significance tests.
3. failure to use non-random selection procedures.
4. failure to adequately control the teacher variable.
5. not defining key terms used in the study.
6. employing measuring instruments of questionable validity and reliability.
7. the use of indiscriminate pooling of data from different groups.
8. failure to use appropriate experimental units of analysis.
9. using multiple t-tests for detecting differences between three or more groups.
10. failure to use multivariate tests for studies employing more than one criterion variable.

11. misuse of ANCOVA when a covariate is influenced by the treatment.
12. failure to control the Hawthorne and novelty effects.

Steps for controlling, and hopefully eliminating, these problems will be discussed in this chapter along with other suggestions for improving the quality of experimental research in education. Solutions for the above problems will be discussed under three main headings: experimental designs, statistical analysis, and research reports.

## Experimental Designs

An experimental research design is to an experimenter as a blueprint is to an architect. A blueprint for a project is generally constructed in such a manner that insures the completion of the project once executed. In other words, the blueprint is designed around the availability of construction materials and technology for carrying out the project. This should also be the case with experimental research designs; they should be capable of execution, and their results should be generalizable. For this later purpose, valid inferential models or statistical tools should be available. Thus, one chooses a design that is capable of valid interpretations.

## Hawthorne and Novelty Effects

Every effort should be made to control the Hawthorne and novelty effects. Many experimenters feel that these effects will be balanced across treatments, and, as a result, will have no differential effects on the dependent variables. But, one cannot be sure of this; these effects may interact with different treatments in different ways. Students should not be told that they are participating in an experiment, and every effort should be taken to insure that the experimental setting approximates a typical classroom setting. If programed materials or unfamiliar techniques, such as discovery treatments, are to be used in the experimental setting, the subjects should have had prior experience with such modes and techniques. In the majority of comparative research studies, the discovery groups were at a disadvantage; these groups were generally unfamiliar with discovery techniques. This unfamiliarity may have been sufficient to render the discovery techniques less than effective in many instances.

One way to control the novelty effects of experimental tasks is to extend the experiment over time. This, in effect, should null out or minimize any disruptive or novelty effects. Then, too, there is always the possibility that these effects will have conditioned the learner and interact with the method of presentation to produce spurious results. This interaction could be experimentally studied.

Many of the comparative research studies have been of short term duration, from one to three days. Generally, these experiments involved novel learning tasks. Both the expository and discovery treatment groups have been influenced by the novelty effects of the learning tasks, and, in addition, the discovery groups were confronted with the disruptive effects of their treatments. The influence of both of these variables could have been minimized by extending the duration of the experiment. Measures could be taken at specified time intervals to estimate novelty and disruptive effects, as well as treatment effects. By employing large numbers of groups, all groups need not be tested at each interval, thus reducing the sensitizing or learning effects of measurement.

## Control of the Teacher Variable

In teaching-methodology studies where the teaching variable is of interest, the treatment variable should be controlled, and some evidence should be presented to indicate that the teacher was faithful to his assigned method. If programed instructional materials are employed, the teacher variable is controlled. But, studies employing programed materials do not necessarily have direct application to the classroom where programed materials are not employed. Since the majority of classroom teaching does not involve programed materials, it appears desirable to conduct teaching experiments using human teachers, as opposed to programed booklets (teachers).

The teacher variable, with non-programed instruction, can be controlled by adequately choosing and training teachers for the study. The training program should include instruction in general subject matter, use of specific methods of instruction, and the procedures for administering and scoring classroom tests. To control the dimension of teacher personality, each teacher could be assigned to teach an expository class and a discovery class.

In order to measure teacher fidelity to treatments, non-student observer ratings and student perception rating scales can be used. The observer ratings may be made by classroom observations, audio recordings, or video recordings. If possible, reliability estimates should be computed for the rating scales. The reliability of student ratings of teacher behavior should be a function of exposure time.

Worthen (1965) used both non-student observer ratings and pupil perception ratings of teaching behavior on the discovery-expository dimension. The later instrument was a forced-choice (ipsitive) questionnaire which elicited pupil responses to statements about teacher-behavior characteristics of their teacher. The questionnaires were scored so that a teacher-index of 100 reflected a pure discovery teaching behavior, while a teacher-index of zero reflected a pure expository teaching behavior. This instrument was administered as both a pretest and a posttest. Pretest scores should reflect the expository model, whereas posttest scores should reflect the experimental treatments. Gains from pretest to posttest teacher index scores can also be used to reflect teaching behavior by observing trends between the two methods of teaching. But, in this case, the reliability of the gain-scores should be taken into consideration; generally, gain-scores are less reliable than either the pretest or posttest scores.

Effort must also be made to control experimental bias. Therefore, it is desirable not to have the experimenter teach all treatment groups (in a small study).

## Retention Measures

Retention is a two-stage process involving storage and retrieval of knowledge. A critical factor in retention is the length of time between storage and retrieval. It is this factor that differentiates between initial learning and retention. As a result, some educators have measured retention in terms of relearning scores; i.e., the time or number of trials to relearn. But, this relearning measure is confounded with learning itself.

The most common measure of retention is recall. Recall is defined as the score on the first relearning measure. Recall is relatively unrelated to learning, assuming that no practice or relearning occurs between acquisition and retrieval. Thus, recall is a function of retention whereas relearning is not.

Typically, an initial-achievement test consists of items for evaluating the extent to which the intended behaviors have been attained. These test items can be at any of Bloom's six levels of cognitive behavior. Thus, the retention measure should be an equivalent form (or the same form) given some time interval following the initial-achievement test. Following this convention, the items on a retention test need not be identical to those items on the initial achievement test; they need only sample the same intended behaviors as the initial test. If the same form of the initial measure is used for the retention measure, an appropriate time interval should elapse between applications of the test to control for the carryover effects of memory of test items or learning from taking the test.

Retention- and initial-learning scores seldom indicate how much or what a student has learned or retained. Such scores usually result from norm-referenced tests which indicate only a person's relative standing with respect to some norming group. Norm-referenced tests are designed to increase the variability among individual scores. As a result, items that are too easy or too difficult are not included since they do not discriminate between individuals. A resulting test score is, therefore, not indicative of what a student knows or has retained.

Criterion-referenced tests, as opposed to norm-referenced tests, are concerned with individual progress. The items on a criterion-referenced test are accurate reflections of the criterion behavior, and are not constructed to differentiate between individuals. Care is taken to identify each test item with a domain of expected behavior; and, as a result, such tests can be used to accurately determine what a person has or has not learned or retained.

Criterion-referenced retention tests offer certain advantages over norm-referenced retention tests. First, they can be used to determine what knowledge is retained and what learned knowledge is forgotten. Further, time intervals can be established over which the criterion is retained. As with a norm-referenced test, a criterion-referenced test can be used to compare treatments or teaching methods.

If several retention tests are to be administered, at different times, care should be taken to control the effects of one test on another. This can be accomplished by randomly assigning each treatment group to subgroups, one subgroup for each retention measure. Each subgroup is then randomly assigned to take one, and only one, retention test.

Reliability estimates should be computed for all norm-referenced retention measures. One should not expect retention measures to be stable over time; therefore, test-retest or stability estimates may not be appropriate in this case.

For criterion-referenced tests, reliability estimates are generally not available. The classical reliability estimates are not applicable for criterion-referenced tests since they are based on the variability of scores. For criterion-referenced tests, variability is irrelevant. It is obvious that a criterion-referenced test should be internally consistent since the items are all tied to a criterion and are similar in nature. Concerning the internal consistency of criterion-referenced tests, Popham and Husek (1969) state:

> . . . But although it may be obvious that a criterion-referenced test should be internally consistent, it is not obvious how to assess the internal consistency. The classical procedures are not appropriate. This is true because they are dependent on score variability. A criterion-referenced test should not be faulted if, when administered after instruction, everyone obtained a perfect score. Yet, that would lead to a zero internal consistency estimate, something measurement books don't recommend. . . . In fact, even stranger things can happen in practice. It is possible for a criterion-referenced test to have a negative internal consistency index and still be a good test. (p. 5)

## Transfer

Ferguson (1956) provided a useful model of transfer which is assumed to be invariant among individuals. His transfer model, in its simplest form, is a mathematical function of three variables, $f(x, t_x, t_y)$. If $y = f(x, t_x, t_y)$, y represents the measurement of performance of some task ($T_2$), x is a measurement of performance on another task ($T_1$), $t_x$ is the amount of practice on $T_1$, and $t_y$

represents the amount of practice on $T_2$. If $T_1 = T_2$, then $t_x = t_y$ and $x = y$, and y is a function of one variable, $t_y$, i.e., $y = g(t_y)$. The graph of g is the usual learning curve. Thus, Ferguson's model suggests that learning is a special case of transfer. If $t_y = 0$ in Ferguson's model, $y = h(x,t_x)$. Here, y is the measurement of the effect of $t_x$ practice on $T_1$ on an unpracticed task $T_2$. The model h suggests that transfer involves learning how to learn. The model h is the typical model employed in most methodology studies comparing discovery and expository methods on transfer measures.

Ferguson's model further enables one to distinguish certain types of transfer. If $h(x,t_x)$ is greater than zero, then the $t_x$ practice on $T_1$ results in positive transfer to $T_2$; whereas, if $h(x,t_x)$ is less than zero, negative transfer results. If $h(x,t_x) = 0$, then the $t_x$ training on $T_1$ has no effect on learning $T_2$. If $T_1$ and $T_2$ are tasks within the same discipline curriculum, then the associated transfer is called vertical transfer. If $T_1$ and $T_2$ are located in separate discipline curriculums, the associated transfer is called horizontal transfer.

In any transfer study in education, it should be specified what training is hypothesized to be transferred

($T_1$) and what the object of the transfer is ($T_2$). Criterion-referenced tests could prove useful for measuring x and $h(x,t_x)$. Bloom's taxonomy of educational objectives in the cognitive domain should also prove useful for evaluation of learned behavior at the various levels.

Two general types of transfer measures (scores) can be identified. One type is exemplified by the frequency of correct responses or by the amount of performance within a given time interval. Such scores increase with learning. If a control group is employed in the design, baseline comparisons can be made by comparing the treatment and the control groups on the transfer criterion. As a result, positive and negative transfer effects can be identified. The other type of transfer measures are exemplified by the number of errors, the time of response, and the number of trials to criterion. Such scores decrease with improvement in performance on a certain task.

The various quantitative expressions for measuring transfer, because of their variety, do not enable one to compare resulting amounts of transfer under different experimental conditions, within or between studies, in any standard or systematic way. As a result, Gagne *et al*.

(1948, p. 122) recommend the use of the following index for measuring transfer:

$$Z = (E - C)/(Q - C)$$

where E is the score made by a treatment group on the transfer criterion, C is the control group's score on the transfer criterion, and Q is the total possible score on the transfer criterion test. As can be seen, Z is unbounded; hence, it can be used to assess both positive and negative transfer. The values of Z are also independent of variations in learning scores. Gagne _et_ _al_. indicate that the values of Z

> . . . relate the transfer attained at different stages in the learning process, and in different learning tasks, to the total improvement possible in each case. (p. 122)

The positive or negative effects of transfer can also be detected by noting any trends in the data that exist from repeated applications of transfer measures. Such a method would be useful for detecting the effects of practice or unrelated learning on transfer performance.

As with repeated-retention measures, care should be taken to control relevant learning between testing periods. Groups could be partitioned so that each student was exposed to only one transfer measure.

Reliability estimates should be computed and reported for all norm-referenced transfer measures.

## Statistical Analysis

Suggestions for improving the data analysis of experimental research will be discussed in this section. Particular attention will be directed to the following areas: type I and type II errors, indiscriminant pooling of the data, multivariate methods, ANCOVA, the experimental unit, multiple comparisons of means, and the strength of effects.

### Type I and Type II Errors

Prior to collecting any experimental data, the probability of rejecting the null hypothesis when true (alpha level) and the probability of rejecting the null hypothesis when false (power level) should both be specified or computed for each significance test of a null hypothesis. Brewer (1972) suggests three reasons why the computation of power is important:

> . . . (1) Such computations can lead the researcher to the conclusion that there is no point in running the study unless the sample size is materially increased; (2) The computation is essential to the interpretation of negative results, that is, failures to reject the null hypothesis; and (3) Computed power gives the researcher an indication of the level of the probability of a valid rejection of the null hypothesis. (p. 391)

According to Cohen (1965, p. 96), if the alternate hypothesis is stated in exact form ($H_1$: $U_A - U_B = c$), any statistical test contains four parameters with three degrees of freedom:

the power of the test, the alpha level, the sample size, and the effect size (c) in the population. Thus, the power is determined once the latter three parameters are specified, or the sample size can be varied to produce the desired degree of power for a fixed alpha level and for a fixed effect size c. The problem for many experimenters is deciding on the size of the population effect. To provide some guidelines, Cohen (1962, 1969) has operationally defined three levels of effects, small, medium, and large. The small difference in population means is defined as .25sd (sd is the population standard deviation), a medium difference as .5sd, and a large difference as sd. The larger the difference, the greater the power. Cohen further recommends a .80 power convention, partially because the consequences of type I errors are more serious than the consequences of type II errors in educational studies. The medium-effect size is recommended whenever in doubt as to the effect size to be used.

With the publication of Cohen's book (Cohen, 1969), the calculation of the power of significance tests becomes a relatively simple matter. His book contains power levels for various tests and combinations of effect and sample sizes. All that is involved in calculating power levels for univariate tests using his tables is arriving at some decision concerning the effect size to be used.

## Pooling of the Data

For many reasons, pooling of data without some justification is not a desirable practice to follow in experimental studies. Students taught in different rooms (under the same treatment), at different times, or by different teachers should not be considered to be under the same experimental conditions. Data from dependent measures for these groups should not be pooled or combined without checking for homogeneity of means and variances for the various groups. But, when this is done, the experimentwise error rate increases. For this reason, one should be discrete when following such a practice.

If an experiment involves an AXB factorial design with a levels under factor A and b levels under factor B, e.g., a treatments-by-levels design, then ab separate treatment groups are involved. As a result, the treatments should be applied to ab separate groups and the data from these groups analyzed as separate units. On some occasions, the treatments are applied simultaneously to all students from the ab treatment groups with a single group at the same time. Such would be the case if programed instructional materials were involved. Here, the different methods are randomly mixed within the group, and care is taken to randomize such variables as seating and lighting within the room. In such controlled situations using individualized instruction, the data can be partitioned into ab data groups and analyzed as such.

Multivariate Methods

Whenever an experiment involves more than one dependent variable, and unless one can determine that the variables are uncorrelated, multivariate statistical methods should be employed. Under certain conditions (dependent measures randomly assigned and compound symmetry of the covariance matrix), repeated-measure designs may be employed for multi-dependent-variable experiments (see Winer, p. 369). Multivariate statistical techniques enable the experimenter to take into consideration the correlation among the dependent variables. Further, exact alpha levels can be obtained from known sampling distributions. Bartlett's sphericity test (Cooley and Lohnes, 1971; p. 103) can be employed to determine whether the dependent variables are significantly correlated to justify using multivariate procedures. If Bartlett's test is not significant, separate univariate tests may be employed to analyze the data on the dependent variables. Bock and Haggard (1968), Cooley and Lohnes (1971), and Tatsuoka (1971) provide excellent descriptions of multivariate techniques and procedures.

## ANCOVA

When using ANCOVA, a covariate should not be influenced by or correlated with the treatment. The reason for this, according to Evans and Anastasio (1968), is that the influence of the treatment on the covariate

> . . . usually produces a linear correlation between the treatment effect and the covariate, with the result that the sum of squares for treatment and the sum of squares associated with the covariate are confounded. (p. 227)

Such would be the case when employing an immediate-posttest measure as covariate for a delayed-posttest measure; both variables have been influenced by the treatment.

## The Experimental Unit

Whenever individualized instruction is not employed in methodology studies, the class mean should be employed as the unit of analysis. Concerning this, Raths (1967) stated:

> It is hoped that graduate student advisors, editors of research journals, and consumers of research will hold in disrepute all studies which attempt to use individual students as the unit of analysis in a methodology study not involving a treatment presented to individuals. (p. 265)

When one uses group means rather than individual subjects as the experimental unit, fewer observations result in a loss of power to detect differences. Concerning this, Peckham _et_ _al_. (1969) state:

> . . . The loss in power is not as great as it appears initially, since variation among class means is much less than among individual pupils within classes. . . . Although the degrees of freedom for the error term are drastically reduced, the corresponding decrease in the error term for the F-test partially compensates for this loss. (p. 344)

Procedures are available for providing evidence concerning _a_ _priori_ assumptions of independent responses of students within classes (see Peckham _et_ _al_., 1969; pp. 345-46).

Multivariate designs (e.g., Hotelling's $T^2$, Wilk's Lambda, or MANOVA) may be used when the class mean is used as the experimental unit. Each classroom can be treated as a single subject, and then treat class components, e.g., sex, I.Q., achievement, retention, transfer, etc., as though they represent measurements on the same subject made under some experimental conditions. The resulting design could also be considered as a two-factor design with repeated measures on the second factor (see Winer, p. 298).

### Multiple Comparisons of Means

Two general multiple comparison methods exist: _a_ _priori_ comparisons and _post_ _hoc_ comparisons. _A_ _priori_ comparisons, e.g., orthogonal contrasts, Dunnett's method, and Bonferroni t-statistics, are used instead of ANOVA. *Post hoc* comparisons, e.g., Scheffe's method, Tukey's method, and the Newman-Keuls test, are used after getting global ANOVA significance. Under no circumstances should *post hoc* procedures be carried out following a non-significant F-test; spurious results may be generated.

The area of multiple comparisons of means appears to be one of the more confusing areas of statistics. Petrinovich and Hardyck (1969, p. 47) note this confusion:

> Textbook authors--at least in the area of psychological statistics--have not been particularly helpful. Authors such as Edwards (1960), Federer (1955), Hays (1963), McNemar (1962), and Winer (1962) either offer no evaluation as to which method is preferable, or preface their remarks with a cautionary statement to the effect that these methods are still under study and that mathematical statisticians are not entirely in agreement concerning the preferred method. Similarly, disagreement exists as to when these methods may be used. Some discussions state that a significant F ratio over all conditions must be obtained before multiple comparison methods can be used (Hays, 1963; McNemar, 1962); other discussions make no mention of such a requirement (Federer, 1955; Winer, 1962) or deny that it is necessary at all (Edwards, 1960; Ryan, 1959a).

Games (1971) provides a clear discussion and summary of multiple-comparison methods. He also offers guidance for using the various methods of multiple comparisons.

## Strength of Effect Measurements

During the last decade considerable attention has been focused on determining the size of experimental effects (degree of association between variables) as well as their statistical significance. The size of an effect is measured as the variance of the dependent variable attributable to the independent variable. As an example, suppose the experimenter effect is indexed by a product-moment correlation coefficient $r$. With an $n$ of 42, an $r$ of .40 is significant at the .05 level (two tails). But $r^2 = (.40)^2 = .16$, which indicates that the independent variable accounts for only 16 per cent of the variance in the dependent variable. An investigator who only reports significant test values (i.e., $F = 4.3$, $t = 3.1$, etc.) may be yielding misleading information. Vaughan and Corballis (1969) have commented on the usefulness and estimates of strength of effects in ANOVA designs.

> Estimates of individual variance components indicates the magnitude, as distinct from the statistical significance, of the variation due to particular effects or interactions. Such estimates may serve a number of useful purposes. A knowledge of the relative contributions of components in a given experiment could guide the researcher in choosing from among a number of specific designs in a subsequent experiment. Again, it may be possible to compare absolute variance estimates between given experiments which employ the same units of measurement. (p. 204)

To this author's knowledge, the experimenters in the area of discovery have not availed themselves of this statistical tool, judged by the absence of effect estimates from the research reports.

Vaughan and Corballis (1969) provide computational formulas appropriate for estimating the strength of effects in basic one-way, two-way, and three-way ANOVA designs. Tatsuoka (1970) provides an estimate for the strength of effects of multivariate ANOVA tests.

## Research Reports

### Describing the Experiment

The results of any experimental research should be reported in such a manner that the experiment could be replicated to some degree by another experimenter if desired. This includes accurately describing the experimental treatments and the conditions under which the experiment took place, e.g., time of day, a description of the learning tasks, a description of the subjects and how they were selected, length of the experiment, and the nature of pupil or teacher involvement. Key terms, such as *retention* and *transfer*, should be defined, and the instruments used to assess these and other constructs should be clearly described, including validity and reliability information.

Whenever feasible, raw data should be included in research reports. If space does not permit publishing of the raw data, it should be reported where and how the raw data can be obtained. If upon investigation, it is found that the data has been improperly analyzed, the data can be reanalyzed using appropriate techniques. Thus, assuming a well-controlled design, the study continues to contribute in a positive way to educational knowledge.

Suydam (1968) constructed an instrument for evaluating experimental research studies. Her instrument consists of nine general categories:

1. How practically or theoretically significant is the problem?
2. How clearly defined is the problem?
3. How well does the design answer the research questions?
4. How adequate does the design control variables?
5. How properly is the sample selected for the design and purpose of the research?
6. How valid and reliabile are the measuring instruments or observational techniques?
7. How valid are the techniques of analysis of the data?
8. How appropriate are the interpretations and generalizations from the data?

9. How adequately is the research reported?

Each question is to be rated on a five-point scale, ranging from poor to excellent. This instrument should be of great assistance in helping the researcher design his experiment. Also, it should prove to be of valuable guidance in aiding the experimenter write his final report of his research, and making sure of the completeness of vital information.

In addition to the information required on Suydam's instrument, power levels for all significance tests (univariate) should be reported. If the effect size is not known, this author recommends the use of a medium effect size of .5 sd.

# CHAPTER VI

# PROBLEMS AND HYPOTHESES FOR FUTURE RESEARCH

## Data Analyses

This study has identified many experiments, whose experimental designs or data analyses could be improved upon, that could, and possibly should, be replicated. For example, the studies by Worthen (1965), Scott (1970), Cooke (1971), Murdoch (1971), and Olander and Robertson (1973) are all worthy of replication.

An interesting research project would be to reanalyze, using appropriate statistical techniques, a subgroup of those comparative (discovery versus expository) research studies that report raw data and contain questionable data analyses. This project should shed light on the types of spurious results that are possible under inappropriate data analyses. Also, the results of such an endeavor may suggest treatment differences along particular dimensions.

## Learning

It appears that the teacher's personality is an important variable in learning. Research is needed along this dimension. Naturalistic research methods lend themselves to this task. By observing teachers teach and interact with students, teacher-personality traits and techniques may be detected that suggest effectiveness with certain types of students. Furthermore, students' performance on a discovery task may differ significantly when their teachers are autocratic as opposed to permissive.

It may well be that the effects of teaching, especially the effects of discovery teaching, may not manifest themselves for years. It should be possible to identify a group of teachers that taught using discovery methods five or ten years ago. A follow-up questionnaire with some of their students may suggest advantages and disadvantages for such teaching strategies. It may also be possible to identify certain personality characteristics that are common to teachers who are successful users of discovery techniques.

Enough research has been conducted using arbitrary or novel learning tasks. What remains to be done is to identify difficult (difficult for the students) subject matter and compare different strategies (e.g., inductive, deductive, expository, discovery, and combinations of the latter) for the most effective learning. This type of research should have direct implications for the classroom.

The amount of practice on discovery learning tasks should have an effect on transfer tasks and retention. This problem needs systematic investigation (see the studies by Gagne and Brown, 1961; Eldredge, 1965; and Meconi, 1967).

The effects of student verbalization following discovery needs further, systematic investigating. It may be advantageous, both to the teacher and student, for the student to verbalize while involved with a discovery task. Gagne and Smith (1962) report some evidence to suggest that such is the case. Also, by verbalizing during a discovery task, the teacher may be able to identify certain learning strategies for certain learning tasks.

It may be possible to design a testing procedure, using selected discovery items of varied difficulties, to distinguish different pupil strategies of discovery. Such a testing situation might involve student interviews.

It appears that the meaningful factor of the subject material may be more salient in learning than the discovery variable. If the material is equally meaningful for both discovery- and expository-treatment groups, there may be no significant differences on achievement, retention, and transfer measures. A major obstacle for such a study is the assessment of meaningfulness.

Transfer studies are needed to assess the transfer of learning across subject areas. These studies should contain control groups so that both positive and negative transfer is capable of being measured.

Search and other problem solving strategies, if learned, should be capable of being transferred to other problem situations. But, there is no guarantee that the problem will be solved once the transfer has taken place. This problem needs investigation. Certain types of strategies, such as searching for patterns, may be more powerful for solving certain types of problems than others. Tuckman et al. (1968) report some research evidence to suggest that limited educational exposure to elegant thinking and problem-solving approaches may induce students to adopt the strategy to search when confronted by transfer situations, but leave them lacking the skill to successfully apply the strategy.

## Teaching

Different discovery strategies (e.g., inductive and deductive) varying the number and kinds (e.g., positive and negative) of examples should be investigated for teaching certain concepts (e.g., algebraic, geometric, conjunctive, etc.) in a systematic fashion. A similar type of research could also be conducted for teaching certain principles. In this case, the number of instances of a rule might be varied. The effects of over-learning might also be investigated.

# LIST OF REFERENCES

Anastasiow, Nicholas, Sally A. Sibley, and Teresa M. Leonhardt. "A Comparison of Guided Discovery, Discovery, and Didactic Teaching of Mathematics to Kindergarten Poverty Children," *American Educational Research Journal*, 7(4): 493-509, November, 1970.

Anderson, G. L. "Quantitative Thinking as Developed Under Connectionist and Field Theories of Learning," In Swenson, Esther J. *et al.*, *Learning Theory in School Situations*. Minneapolis: University of Minnesota Press, 1949.

Ashton, Sister Madeleine Rose. "Heuristic Methods in Problem Solving in Ninth-Grade Algebra," Unpublished Ph.D. thesis, Stanford University, 1962.

Ausubel, D. P. "Learning by Discovery: Rationale and Mystique," *National Association of Secondary School Principles* (Bulletin), 45: 18-58, December, 1961.

Ausubel, D. P. *The Psychology of Meaningful Learning*. New York: Greene and Stratton, 1963.

Ausubel, D. P. and Floyd G. Robinson. *School Learning: An Introduction to Educational Psychology*. New York: Holt, Rinehart, and Winston, 1969. (Chapter 16, pp. 478-503)

Babikian, Yeghia. "An Empirical Investigation to Determine the Relative Effectiveness of Discovery, Laboratory, and Expository Methods of Teaching Science Concepts," *Journal of Research in Science Teaching*, 8(3): 201-09, 1971.

Bagley, William C. *The Educative Process*. New York: Macmillan, 1905.

Bakan, David. "The Test of Significance in Psychological Research," Psychological Bulletin, 66(6): 423-37, December, 1966.

Ballew, H. "Discovery Learning and Critical Thinking in Algebra," High School Journal, 50: 261-70, 1967.

Barrish, Bernard. "Inductive Versus Deductive Teaching Strategies With High and Low Divergent Thinkers," Unpublished Ed.D. thesis, Stanford University, 1970.

Bassler, Otto C., Warren H. Hill Jr., Josephine A. Ingle, and Billie Earl Sparks. "Comparison of Two Levels of Guidance in Teaching Elementary School Mathematics," School Science and Mathematics, 71: 303-12, 1971.

Beberman, Max. An Emerging Program of Secondary School Mathematics. Cambridge, Massachusetts: Harvard University Press, 1958.

Belcastro, Frank P. "Relative Effectiveness of the Inductive and Deductive Methods of Programing Algebra," The Journal of Experimental Education, 34: 77-82, Spring 1966.

Bittinger, M. L. "Review of Discovery," The Mathematics Teacher, 61: 140-46, February 1968.

Bock, R. D. and E. A. Haggard. "The Use of Multivariate Analysis of Variance in Behavioral Research," In Whitla, Dean K. (Ed.), Handbook of Measurement and Assessment in Behavioral Sciences, Reading, Massachusetts: Addison-Wesley, 1968. (pp. 100-142).

Boeck, C. H. "The Inductive-Deductive Compared to the Deductive-Descriptive Approach to Laboratory Instruction in High School Chemistry," Journal of Experimental Education, 19: 247-53, March 1951.

Box, G. E. P. "Non-normality and Tests on Variance," Biometrika, 40: 318-35, 1953.

Bracht, Glenn H. and Gene V. Glass. "The External Validity of Experiments," American Educational Research Journal, 5: 437-74, November 1968.

Brenner, Charles J. "An Experimental Comparison of Direct-Detailed Versus Directed Discovery Laboratory Exercises in Teaching Selected Elements of Basic Electricity," Unpublished Ed.D. thesis, University of Missouri, 1968.

Brewer, James K. "On the Power of Statistical Tests in the American Educational Research Journal," *American Educational Research Journal*, 9(3): 391-401, Summer 1972.

Britton, Gwyneth E. "A Comparison of the Inductive and Deductive Group Approaches in Teaching Selected Phonic Generalizations to Second Grade Children," Unpublished Ed.D. thesis, Oregon State University, 1969.

Brown, Frederick G. *Principles of Educational and Psychological Testing*. Hinsdale, Illinois: The Dreyden Press, Inc., 1970.

Brown, Lynn H. "A Comparison of a 'Teaching For Thinking' Approach and a Conventional Approach to the Teaching of Algebra I," Unpublished Ph.D. thesis, University of Iowa, 1969.

Brudzynski, Alfred J. "A Comparative Study of Two Methods For Teaching Electricity and Magnetism With Fifth and Sixth Grade Children," Unpublished Ed.D. thesis, Boston University School of Education, 1966.

Bruner, J. S. *The Process of Education*. Cambridge, Massachusetts: Harvard University Press, 1960.

Bruner, J. S. "The Act of Discovery." *Harvard Educational Review*, 31: 21-32, 1961.

Bruner, J. S. *Toward a Theory of Instruction*. Massachusetts: Belknap Press, 1966.

Carroll, John B. "Words, Meanings and Concepts," *Harvard Educational Review*, 34(2): 202, 1964.

Caruso, George E. "A Comparison of Two Methods of Teaching the Mathematical Theory of Groups to College Freshmen," Unpublished Ph.D. thesis, New York University, 1966.

Chambers, D. W. "Putting Down the Discovery Learning Hypothesis," *Educational Technology*, 11: 54-9, March 1971.

Cochran, W. G. "Errors of Measurement in Statistics," *Technometrics*, 10(4): 637-666, November 1968.

Cohen, Jacob. "The Statistical Power of Abnormal-Social Psychological Research," *Journal of Abnormal and Social Psychology*, 65: 145-53, 1962.

Cohen, Jacob. "Some Statistical Issues in Psychological Research," In B. B. Wolman (Ed.). *Handbook of Clinical Psychology*. New York: McGraw-Hill, 1965. (pp. 95-121).

Cohen, Jacob. *Statistical Power Analysis for the Behavioral Sciences*. New York: Academic Press, 1969.

Colburn, Warren. *Intellectual Arithmetic, Upon the Inductive Method of Instruction*. Boston: Hilliard, Grey, Little, and Wilkins, 1828.

Cook, Desmond L. *The Impact of the Hawthorne Effect in Experimental Designs in Educational Research*. Cooperative Research Project, No. 1757, U. S. Office of Education, June 1967.

Cooke, Gary E. "Conceptual Learning in Young Children: A Comparison of the Effects of Rote, Principle, and Guided Discovery Strategies on Conceptualization in First Grade Children," Unpublished Ed.D. thesis, University of Oregon, 1971.

Cooley, W. W. and P. R. Lohnes. *Multivariate Data Analysis*. New York: Wiley, 1971.

Craig, Robert C. "Discovery, Task Completion, and the Assignment as Factors in Motivation," *American Educational Research Journal*, 2(4): 217-22, November 1965.

Craig, Robert C. "Recent Research of Discovery," *Educational Leadership*, 26: 501, Fall 1969.

Craik, Eva Lee. "The Relative Effectiveness of the Inductive-Deductive and the Deductive-Descriptive Methods in the Teaching of College Zoology," Unpublished Ed.D. thesis, North Texas State University, 1966.

Cronbach, Lee J. "The Logic of Experiments on Discovery," In Shulman, L. S. and E. R. Keislar (Eds.), *Learning by Discovery: A Critical Appraisal*, Skokie, Illinois: Rand McNally and Company, 1966. (pp. 76-92).

Davis, R. B. "The Range of Rhetorics, Scale and Other Variables," *Journal of Research and Development in Education*, 1(1): 51-74, Fall 1967.

Denmark, Ewell Thomas. "A Comparative Study of Two Methods of Teaching Elementary Algebra Students to Use the Algebraic Technique to Solve Verbal Problems," Unpublished Ph.D. thesis, The Florida State University, 1964.

Dennison, Clifford C. "Evaluating a Verbal Approach and a Discovery Approach in Learning Selected Science Principles at Two Levels of Maturity," Unpublished Ed.D. thesis, The University of Florida, 1969.

Dixon, Wilfrid J. and Frank J. Massey, Jr. *Introduction to Statistical Analysis*. New York: McGraw-Hill, 1969.

Duncan, C. P. "Learning To Learn In Response-Discovery and in Paired-Associate Lists," *American Journal of Psychology*, 77: 367-79, 1964.

Edwards, A. J. and Dale P. Scannell. *Educational Psychology*. Scranton, Pennsylvania: International Textbook Co., 1968.

Eldredge, Garth M. "Expository and Discovery Learning in Programed Instruction," Unpublished Ph.D. thesis, University of Utah, 1965.

Evans, Selby H. and Ernest J. Anastasio. "Misuse of Analysis of Covariance When Treatment Effect and Covariate Are Confounded," *Psychological Bulletin*, 69: 225-34, April 1968.

Ferguson, G. A. "On Transfer and the Abilities of Man," *Canadian Journal of Psychology*, 10: 121-31, 1956.

Foord, Marion. "Arithmetic: Inductive Versus Deductive Methods of Teaching Area by Programed Instruction," *Educational Review*, 16: 130-37, February 1964.

Fullerton, Craig Kerr. "A Comparison of the Effectiveness of Two Prescribed Methods of Teaching Multiplication of Whole Numbers," Unpublished Ph.D. thesis, State University of Iowa, 1955.

Gabor, Georgia M. "Teaching Methods and Incentives in Relation to Junior High Mathematics Achievement," *California Journal of Educational Research*, 23(2): 56-70, March 1972.

Gagne, R. M., Harriet Foster, and Meriam E. Crowley. "The Measurement of Transfer of Training," *Psychological Bulletin*, pp. 97-130, March 1948.

Gagne, R. M. and L. T. Brown. "Some Factors in the Programming of Conceptual Learning," *Journal of Experimental Psychology*, 62: 313-21, 1961.

Gagne, R. M. and E. Smith, Jr. "A Study of the Effects of Verbalization on Problem Solving," *Journal of Experimental Psychology*, 63: 12-18, 1962.

Gagne, R. M. *The Conditions of Learning*. New York: Holt, Rinehart, and Winston, 1965.

Gagne, R. M. "Varieties of Learning and the Concept of Discovery," In Shulman, L. S. and E. R. Keislar (Eds.) *Learning by Discovery: A Critical Appraisal*. Chicago: Rand McNally, 1966. (pp. 135-150).

Gagne, R. M. *The Conditions of Learning*. New York: Holt, Rinehart, and Winston, 1970.

Games, Paul A. "Multiple Comparisons of Means," *American Educational Research Journal*, 8(3): 531-65, May 1971.

Gaston, Jane A. and John R. Kolb. "A Comparison of Three Strategies For Teaching a Selected Mathematical Concept to Students in College Algebra," *Journal for Research in Mathematics Education*, 4(3): 177-86, May 1973.

Gentry, Castelle G. "Relative Effectiveness of Discovery and Expository Methods of Teaching Concepts Through the Single-Concept Film," Unpublished Ed.D. thesis, Michigan State University, 1965.

Glass, G. V. and J. Stanley. *Statistical Methods in Education and Psychology*. New York: Prentice-Hall, 1969.

Green, Thomas F. *The Activities of Teaching*. New York: McGraw-Hill Book Co., 1971.

Grote, C. N. "A Comparison of the Relative Effectiveness of Direct-Detailed and Directed Discovery Methods of Teaching Selected Principles of Mechanics in the Area of Physics," Unpublished Ed.D. thesis, University of Illinois, 1960.

Hanson, Lawrence E. "Inductive Discovery Learning, Reception Learning, and Formal Verbalization of Mathematical Concepts," Unpublished Ph.D. thesis, The Florida State University, 1967.

Hays, W. L. *Statistics*, New York: Holt, Rinehart, and Winston, 1963.

Henderson, Kenneth B. "A Model for Teaching Mathematical Concepts," *The Mathematics Teacher*, 60: 573-7, 1967.

Hendrickson, G. and W. H. Schroeder. "Transfer of Training and Learning to Hit a Submerged Target," *Journal of Educational Psychology*, 32: 205-13, 1941.

Hendrix, Gertrude. "A New Clue to Transfer of Training," *Elementary School Journal*, 48: 197-208, 1947.

Hermann, G. "Learning By Discovery: A Critical Review of Studies," *The Journal of Experimental Education*, 38: 58-72, Fall 1969.

Hermann, G. "Egrule Versus Ruleg Teaching Methods: Grade, Intelligence, and Category of Learning," *Journal of Experimental Education*, 39: 22-33, Spring 1971.

Hirsch, Christian R. "An Experimental Study Comparing the Effects of Guided Discovery and Individualized Instruction on Initial Learning, Transfer, and Retention of Mathematical Concepts and Generalizations," Unpublished Ph.D. thesis, University of Iowa, 1972.

Hummel, Thomas J. and Joseph R. Sligo. "Empirical Comparison of Univariate and Multivariate Analysis of Variance Procedures," *Psychological Bulletin*, 76(1): 49-57, July 1971.

Hunt, Earl B., Janet Marin, and Philip J. Stone. *Experiments in Induction*. New York: Academic Press, 1966.

Jamieson, G. H. "Learning by Programmed and Guided Discovery Methods at Different Age Levels," *Programmed Learning and Educational Technology*, 6: 26-30, January 1969.

Jamieson, G. H. "Transfer of Learning Under Two Conditions of Instruction: Programed and Guided Discovery," *Programmed Learning and Educational Technology*, 7: 113-19, April 1970.

Jamieson, G. H. "Learning and Retention: A Comparison Between Programmed and Discovery Learning at Two Age Levels," *Programmed Learning and Educational Technology*, 8: 34-40, January 1971.

Johnson, Mauritz. "Who Discovered Discovery?" *Phi Delta Kappan*, 48(2): 120-3, October 1966.

Jones, Phillip S. "Discovery Teaching--From Socrates to Modernity," *The Arithmetic Teacher*, 17: 503-10, October 1970.

Judd, C. H. "The Relation of Special Training to General Intelligence," *Educational Review*, 36: 28-42, 1908.

Kanes, LeLage G. "A Comparison of Two Teaching Strategies Used to Present a Unit in Elementary Mathematics Using Computer-Assisted Instruction," Unpublished Ed.D. thesis, University of Pennsylvania, 1971.

Keese, Earl E. "A Study of the Creative Thinking Ability and Student Achievement in Mathematics Using Discovery and Expository Methods of Teaching," Unpublished Ph.D. thesis, Peabody College for Teachers, 1972.

Kellogg, Theodore E. "The Relative Effects of Variations in Pure and Physical Approaches to the Teaching of Euclidean Geometry on Pupils' Problem Solving Ability," Unpublished Ph.D. thesis, University of Minnesota, 1956.

Kempthorne, Oscar. "The Design and Analysis of Experiments With Some Reference to Educational Research," In Collier, Raymond O. and Stanley M. Elam (Eds.). *Research Design and Analysis: Second Annual Phi Delta Kappa Symposium on Educational Research*, Indiana: Phi Delta Kappa, 1961.

Kersh, B. Y. "The Adequacy of 'Meaning' as an Explanation for Superiority of Learning by Independent Discovery," *Journal of Educational Psychology*, 49: 282-92, October 1958.

Kersh, B. Y. "The Motivating Effect of Learning By Directed Discovery," *Journal of Educational Psychology*, 53: 65-71, 1962.

Kersh, B. Y. "Learning by Discovery: What is Learned?" *The Arithmetic Teacher*, 11: 226-32, April 1964.

Kersh, B. Y. and M. C. Wittrock. "Learning by Discovery: An Interpretation of Recent Research," *Journal of Teacher Education*, 13: 461-68, 1962.

Keurst, Arthur J. and Joanna M. Martin. "Rote Versus Discovery Learning," *School and Community*, pp. 42, 44, November 1968.

Krumboltz, J. D. and W. W. Yabroff. "The Comparative Effects of Inductive and Deductive Sequences in Programed Instruction," *American Educational Research Journal*, 2(4): 223-35, November 1965.

Kuhfittig, Peter K. "The Effectiveness of Discovery Learning in Relation to Concrete and Abstract Teaching in Mathematics," Unpublished Ph.D. thesis, George Peabody College for Teachers, 1972.

Lackner, Lois M. "The Teaching of the Limit and Derivative Concepts in Beginning Calculus by Combinations of Inductive and Deductive Approaches," Unpublished Ph.D. thesis, University of Illinois, 1968.

Lahnston, Anton T. "A Comparison of Directed Discovery and Demonstration Strategies for Teaching Geographic Concepts and Generalizations," Unpublished Ph.D. thesis, University of Washington, 1972.

LaRocque, G. E. "The Effectiveness of the Inductive and Deductive Methods of Teaching Figurative Language to Eighth Grade Students," Unpublished Ph.D. thesis, Stanford University, 1965.

Lennek, David. "Open-ended Experiments in Junior High School Science: A Study of Their Effect on the Acquisition of Science Information, Laboratory Skills, and Attitudes Towards Science," Unpublished Ed.D. thesis, Columbia University, 1967.

Levine, Joan L. "A Comparative Study of Two Methods of Teaching Mathematical Analysis at the College Level," Unpublished Ed.D. thesis, Columbia University, 1967.

Luck, William E. "An Experimental Comparison of Direct and Detailed Method and Directed-Discovery Method of Teaching Selected Automotive Topics to Senior High School Industrial Arts Students," Unpublished Ed.D. thesis, Oklahoma State University, 1966.

May, Lola June. "A Statistical Comparison of the Effectiveness of Teaching Per Cent by the Traditiona, Ratio, and Discovery Methods," Unpublished Ph.D. thesis, Northwestern University, 1965.

Maynard, Freddy J. "A Comparison of Three Methods of Teaching Selected Content in Eighth and Ninth Grade General Mathematics Courses," Unpublished Ed.D. thesis, University of Georgia, 1969.

McConnell, T. R. "Discovery Vs. Authoritative Identification in the Learning of Children," Studies in Education, 9(5): 13-60, 1934.

McMurray, Charles A. and Frank M. The Method of the Recitation. New York: MacMillan, 1897.

Meconi, L. J. "Concept Learning and Retention in Mathematics," The Journal of Experimental Education, 36(1): 51-7, Fall 1967.

Michael, R. E. "The Relative Effectiveness of Two Methods of Teaching Certain Topics in Ninth Grade Algebra," The Mathematics Teacher, 42: 83-7, February 1949.

Moss, J., Jr. "An Experimental Study of the Relative Effectiveness of the Direct-Detailed and the Directed Discovery Methods of Teaching Letterpress Imposition," Unpublished Ed.D. thesis, University of Illinois, 1960.

Murdoch, Robert L. "Concept Learning and Retention: Effects of Presentation Method and Learning Procedure for Different Intellectual Levels," Unpublished Ph.D. thesis, State University of New York at Albany, 1971.

Myers, Jerome L. Fundamentals of Experimental Design. Boston: Allyn and Bacon, 1966.

National Council of Teachers of Mathematics. Fifteenth Yearbook: The Place of Mathematics in Secondary Education. New York: Bureau of Publications, Teachers College, Columbia University, 1940.

Naughton, Sister Marie. "An Experimental Comparison of the Inductive, Deductive, and Thought Methods of Teaching Spelling on the Third-, Fourth-, and Fifth-Grade Levels," Unpublished Ph.D. thesis, Fordham University, 1962.

Nelson, Barbara A. and Dorothy A. Frayer. Discovery Learning Vs. Expository Learning: New Insight Into an Old Controversy. Wisconsin: Research and Development Center for Cognitive Learning, Wisconsin University, April 1972. (ERIC No. ED 061532).

Neuhouser, David Lee. "A Comparison of Three Methods of Teaching a Programmed Unit on Exponents to Eighth Grade Students," Unpublished Ph.D. thesis, The Florida State University, 1964.

Nichols, Edith J. "A Comparison of Two Methods of Instruction in Multiplication and Division for Third-Grade Pupils," Unpublished Ed.D. thesis, University of California, Los Angeles, 1971.

Norman, Martha. "Three Methods of Teaching Basic Division Facts," Unpublished Doctoral thesis, State University of Iowa, 1955.

Olander, Herbert T. and Howard C. Robertson. "The Effectiveness of Discovery and Expository Methods in the Teaching of Fourth-Grade Mathematics," *Journal For Research in Mathematics Education*, 4(1): 33-44, January 1973.

Page, David P. *Theory and Practice of Teaching*. Syracuse: Hull and Dickson, 1847.

Peckham, Percy D., Gene V. Glass, and Kenneth D. Hopkins. "The Experimental Unit in Statistical Analysis," *Journal of Special Education*, 3(4): 337-49, Winter 1969.

Peters, Donald L. "Discovery Learning in Kindergarten Mathematics," *Journal for Research in Mathematics Education*, 1: 76-87, March 1970.

Petrinovich, L. F. and C. D. Hardyck. "Error Rates for Multiple Comparison Methods: Some Evidence Concerning the Frequency of Erroneous Conclusions," *Psychological Bulletin*, 71: 43-54, 1969.

Popham, W. J. and T. R. Nusek. "Implications of Criterion-Referenced Measurement," *Journal of Educational Measurement*, 6(1): 1-9, Spring 1969.

Price, Jack S. "Discovery: Its Effects on the Achievement and Critical Thinking of Tenth Grade General Mathematics Students," Unpublished Ed.D. thesis, Wayne State University, 1965.

Price, Jack S. "Discovery: Its Effect on Critical Thinking and Achievement in Mathematics," *The Mathematics Teacher*, 60: 874-76, December 1967.

Raths, James. "The Appropriate Experimental Unit," *Educational Leadership*, 25: 263-66, December 1967.

Ray, W. E. "An Experimental Comparison of Direct-Detailed and Direct-Discovery Methods for Teaching Micrometer Principles and Skills," Unpublished Ed.D. thesis, University of Illinois, 1957.

Richardson, Verlin and John W. Renner. "A Study of the Inquiry-Discovery Method of Laboratory Instruction, *Journal of Chemical Education*, 47(1): 77-9, January 1970.

Rizzuto, Malcolm F. "Experimental Comparison of Inductive and Deductive Methods of Teaching Concepts of Language Structure," *The Journal of Educational Research*, 63(6): 269-73, February 1970.

Romberg, Thomas A. and M. Vere DeVault. "Mathematics Curriculum: Needed Research," *Journal of Research and Development in Education*, 1(1): 95-110, Fall 1967.

Roscoe, John T. *Fundamental Research Statistics for the Behavioral Sciences*. New York: Holt, Rinehart, and Winston, 1969.

Rowlett, J. D. "An Experimental Comparison of Direct-Detailed and Directed Discovery Methods of Teaching Orthographic Projection Principles and Skills," Unpublished Ed.D. thesis, University of Illinois, 1960.

Rowlett, J. D. "An Experimental Comparison of Direct-Detailed and Directed Discovery Methods of Presenting Tape-Recorded Instruction," Richmond Virginia: Eastern Kentucky State College, Report No. NDEA VII-629, 1964. (ERIC ED 003183).

Ryan, Thomas A. "Multiple Comparisons in Psychological Research," *Psychological Bulletin*, 56(1): 26-47, 1959.

Sakmyster, Diane Carol Decker. "Comparison of Inductive and Deductive Programmed Instruction on Chemical Equilibrium for High School Chemistry Students," Unpublished Ed.D. thesis, Indiana University, 1972.

Schaaf, Oscar F. "Student Discovery of Algebraic Principles as a Means of Developing Ability to Generalize," Unpublished Ph.D. thesis, The Ohio State University, 1954.

Scott, Joseph A. "The Effects on Short and Long Term Retention and on Transfer of Two Methods of Presenting Selected Geometry Concepts," Wisconsin: Wisconsin University Research and Development Center for Cognitive Learning. Report No. TR138, July 1970. (ERIC ED 044314).

Shelton, Ronald M. "A Comparison of Achievement Resulting From Teaching the Limit Concept in Calculus by Two Different Methods," Unpublished Ph.D. thesis, University of Illinois, 1965.

Shulman, L. S. "Psychological Controversies in the Teaching of Mathematics," In Aichele, Douglas B. and Robert E. Reys (Eds.), *Readings in Secondary School Mathematics*, Boston: Prindle, Weber, and Schmidt, 1971. (pp. 178-192).

Sobel, Max A. "A Comparison of Two Methods of Teaching Certain Concepts in Ninth Grade Algebra," Unpublished Ph.D. thesis, Columbia University, 1954.

Spencer, Herbert. *Education: Intellectual, Moral and Physical*. London: Hurst and Co., 1860.

Steck, J. C. "The Independence of Observations Obtained in Classroom Research," Unpublished Master of Arts thesis, University of Maryland, 1966.

Stock, Suzanne Jane Foster. "A Comparison of An Abstract Deductive and a Concrete Inductive Approach to Teaching the Concepts of Limits, Derivatives, and Continuity in a Freshman Calculus Course," Unpublished Ph.D. thesis, The Ohio State University, 1971.

Strickland, James F. "A Comparison of Three Methods of Teaching Selected Content in General Mathematics," Unpublished Ed.D. thesis, University of Georgia, 1968.

Suchman, J. R. "The Elementary School Training Program," Scientific Inquiry, Urbana, Illinois: University of Illinois Press, 1962.

Suydam, Marilyn N. "An Instrument for Evaluating Experimental Educational Research Reports," The Journal of Educational Research, 61(5): 200-03, January 1968.

Swenson, Esther J. "Organization and Generalization as Factors in Learning, Transfer, and Retroactive Inhibition," In Swenson, Esther J. et al. (Eds.), Learning Theory in School Situations, Minnesota: University of Minnesota Press, 1949.

Tanner, R. Thomas. "Expository-Deductive Versus Discovery-Inductive Programming of Physical Science Principles," Journal of Research in Science Teaching, 6(2): 136-42, 1969.

Tatsuoka, Maurice M. "Statistics," Review of Educational Research, 39(5): 739-43, December 1969.

Tatsuoka, Maurice M. Discriminant Analysis: The Study of Group Differences. Champaign, Illinois: Institute for Personality and Ability Testing, 1970.

Tatsuoka, Maurice M. Multivariate Analysis: Techniques for Educational and Psychological Research. New York: John Wiley and Sons, 1971.

Thiele, C. L. The Contribution of Generalization to the Learning of Addition Facts. New York: Teachers College, Columbia University, 1938.

Thorndike, R. L. and Elizabeth Hagen. Measurement and Evaluation in Psychology and Education. New York: Wiley and Sons, 1961.

Tomlinson, R. M. "A Comparison of Four Methods of Presentation for Teaching Complex Technical Material," Unpublished Ed.D. thesis, University of Illinois, 1962.

Tuckman, Bruce W., James Henkelman, Gerald P. O'Shaughnessy, and Mildred B. Cole. "Induction and Transfer of Search Sets," Journal of Educational Psychology, 59: 59-68, 1968.

Twelkner, Paul A. *Two Types of Teacher-Learner Interaction in Learning by Discovery*, Technical Report No. R-59, Monmouth: Oregon State System of Higher Education, September 1967.

Vaughan, G. M. and M. C. Corballis. "Beyond Tests of Significance: Estimating Strength of Effects in Selected ANOVA Designs," *Psychological Bulletin*, 72(3): 204-13, September 1969.

Werdelin, Ingvar. "The Value of External Direction and Individual Discovery in Learning Situations," *Scandinavian Journal of Psychology*, 9: 241-47, 1968.

Wiesner, Carol. "A Comparison of the Effectiveness of Discovery Versus Didactic Methods and Teacher-Guided Versus Independent Procedures in Principle Learning," *Journal of Educational Research*, 64(5): 217-19, January 1971.

Winch, W. H. *Inductive Versus Deductive Methods in Teaching: An Experimental Research*. Baltimore: Warwick and York, 1913.

Winer, B. J. *Statistical Principles in Experimental Design*. New York: McGraw-Hill, 1962.

Wolfe, M. S. "Effects of Expository Instruction in Mathematics on Students Accustomed to Discovery Methods," Unpublished Ph.D. thesis, University of Illinois, 1963.

Woodward, Ernest L. "A Comparative Study of Teaching Strategies Involving Advanced Organizers and Post Organizers and Discovery and Nondiscovery Techniques Where the Instruction is Mediated by the Computer," Unpublished Ed.D. thesis, The Florida State University, 1966.

Worthen, B. R. "A Study of Discovery and Expository Presentation: Implications for Teaching," *The Journal of Teacher Education*, 19(2): 223-42, Summer 1968.

Worthen, B. and James R. Collins. "Reanalysis of Data From Worthen's Study of Sequencing In Task Presentation," *Journal of Educational Psychology*, 62(1): 15-16, 1971.

Yabroff, W. W. "The Comparative Effects of Inductive and Deductive Sequences in Programmed Instruction," Unpublished Ph.D. thesis, Stanford University, 1963.

Young, J. W. A. *The Teaching of Mathematics*. New York: Longmans and Green, 1906.

Zubulake, George R. "A Study of the Learning by Discovery Controversy in Science Teaching," Unpublished Ph.D. thesis, The University of Michigan, 1970.

# VITA

Richard Charles Weimer was born on January 21, 1939, in Meyersdale, Pennsylvania, where he attended elementary, junior and senior high school. The B.S. in education was awarded him at California State College, Pennsylvania, in January 1960. Upon graduation, he taught mathematics at Canonsburg Junior High School, Canonsburg, Pennsylvania, for one semester. The following three years found him teaching high school mathematics at Shanksville, Pennsylvania. He was a member of the National Science Foundation Academic Year Institute at the University of Illinois during the academic period 1963-64, receiving his A.M. degree in mathematics in 1964. In the fall of 1964, he joined the mathematics staff at Edinboro State College, Pennsylvania, as an instructor of mathematics, and remained there for a period of two years. Since that time, he has been employed by Frostburg State College, Maryland, presently holding the rank of associate professor of mathematics. In 1969, he was awarded a Science Faculty Fellowship to attend the University of Illinois to study mathematics for a period of

15 months. During the 1973-74 academic year, he was on a sabbatical leave from Frostburg State College, and worked half-time at the University of Illinois as a field supervisor for student teachers in the area of mathematics.